

**R**ORSCHACH  
SCIENCE

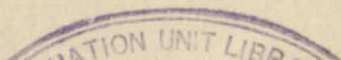
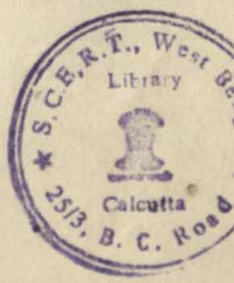
*Rorschach Science*



RORSCHACH  
SCIENCE · *Readings*  
*in Theory & Method*

*Edited by* MICHAEL HIRT

THE FREE PRESS OF GLENCOE



Copyright © 1962 by THE FREE PRESS OF GLENCOE,  
A Division of The Macmillan Company  
Printed in the United States of America

All rights in this book are reserved. No part of this book  
may be used or reproduced in any manner whatsoever  
without written permission except in the case of brief  
quotations embodied in critical articles and reviews.

For information, address:  
The Free Press of Glencoe  
A Division of The Macmillan Company,  
The Crowell-Collier Publishing Company  
60 Fifth Avenue, New York 11

DESIGNED BY MINA BAYLIS

3151  
T., West Bengal

4-3-85

3051

Library of Congress Catalog Card Number: 62-15343

*Key*

137.842

HIR

EVALUATION UNIT LIBRARY	
DAVID HARE TRAINING COLLEGE. CALCUTTA-19	
Acc.No	0033
Date	30 MAR 1965
Call	137.842
No.	4671

*65/80*

## PREFACE

THIS BOOK offers a collection of papers on problems directly related to the development of a rationale for the application of the Rorschach test under conditions that will maximize its validity. Its major objectives are (1) to encourage a more critical evaluation of the Rorschach and its various clinical uses, and (2) to orient the advanced student toward the more fundamental problems of behavior measurement by means of techniques such as the Rorschach. It is not intended to teach methods of either administration or interpretation but rather to help clarify the nature of the pitfalls in the present problems of Rorschach examination.

These objectives require little justification. Although many psychologists have become increasingly aware of the need to modify existing Rorschach theory and practice, there does not appear to have been a systematic effort in this direction. The need for the development of Rorschach principles that are firmly rooted in test and measurement theory is being increasingly recognized.

It is hoped that this book will fill a gap in available textbooks. At the present time no book adequately treats the Rorschach apart from relatively specific clinical applications. The stimulus for preparation of this volume was the recognition of the need for such a textbook while I was a graduate student learning about the Rorschach and a practicing clinician trying to utilize it.

A few words of explanation as to the criteria of selecting the articles



included in this volume are in order. The basic consideration was that an article not be limited to a narrow clinical use. I sought articles which attempted to evaluate the Rorschach as a measuring instrument, not as a predictor of specific clinical behavior. For those selections that are directly based on experimental data, the major criterion was the adequacy of their methodology.

I should like to acknowledge the kindness of the various authors, editors, and publishers who gave permission to reproduce their work. Special thanks are due to Drs. Richard A. Cook, James Layman, Bernard L. Mooney, and J. M. Sadnavitch for their encouragement and critical evaluation during the preparation of this book. My first teachers in the Rorschach, Drs. Don W. Dysinger and Marshall R. Jones, were particularly instrumental in imbuing me with the questioning attitude underlying this book. Finally, I wish to express my gratitude to my wife, who bore graciously my various reactions to the numerous frustrations and irritations accompanying the preparation of any manuscript.

MICHAEL HIRT

# CONTENTS

*Preface*

v

## I

### *Introduction*

#### MEASUREMENT AND THE RORSCHACH

Michael Hirt

3

#### PERSONALITY ASSESSMENT AND PERCEPTION

Bernard Mooney

17

## II

### *The Projective Method*

#### PROJECTIVE METHODS FOR THE STUDY OF PERSONALITY

Lawrence K. Frank

31

#### PROJECTIVE METHODS

*Their Origins, Theory, and Application in Personality Research*

Helen Sargent

53

#### EXPERIMENTALLY INDUCED VARIATIONS IN RORSCHACH

PERFORMANCE Edith Lord

101

### III

#### Scoring

<i>STATISTICAL TESTS OF CERTAIN RORSCHACH ASSUMPTIONS</i>	
<i>Analyses of Discrete Responses</i> J. R. Wittenborn	147
<i>A FACTOR ANALYSIS OF RORSCHACH SCORING CATEGORIES</i>	
J. R. Wittenborn	163
<i>STATISTICAL TESTS OF CERTAIN RORSCHACH ASSUMPTIONS</i>	
<i>The Internal Consistency of Scoring Categories</i> J. R. Wittenborn	176
<i>RORSCHACH SUMMARY SCORES IN DIFFERENTIAL DIAGNOSIS</i>	
Irwin J. Knopf	202

### IV

#### Validity

<i>RORSCHACH VALIDATION</i>	
<i>Some Methodological Aspects</i> Leonard I. Schneider	215
<i>A DUAL APPROACH TO RORSCHACH VALIDATION</i>	
<i>A Methodological Study</i> James O. Palmer	232
<i>CONSTRUCT VALIDITY IN PSYCHOLOGICAL TESTS</i>	
Lee J. Cronbach and Paul E. Meehl	264

### V

#### Reliability

<i>THE RELIABILITY OF THE RORSCHACH INK-BLOT TEST</i>	
Marguerite R. Hertz	295
<i>RORSCHACH SCORER RELIABILITY</i>	
Richard H. Dana	310
<i>INCREMENTS AND CONSISTENCY OF PERFORMANCE IN FOUR</i>	
<i>REPEATED RORSCHACH ADMINISTRATIONS</i> Bert Kaplan and Stanley Berger	314

### VI

#### Current Status

<i>THE RORSCHACH IN PSYCHOLOGICAL SCIENCE</i>	
L. L. Thurstone	325



*FAILURES OF THE RORSCHACH TECHNIQUE*

Joseph Zubin

331

*STATISTICAL METHODS APPLIED TO RORSCHACH SCORES*

*A Review* Lee J. Cronbach

347

VII

*Summary*

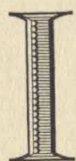
*CURRENT PROBLEMS IN RORSCHACH THEORY AND*

*TECHNIQUE* Marguerite R. Hertz

391

Indexes

431



## *Introduction*

*Michael Hirt*

# MEASUREMENT AND THE RORSCHACH

## PROBLEMS OF PREDICTION

PSYCHOLOGISTS, PARTICULARLY those working in the clinical areas, have been increasingly concerned with the prediction of behavior. To this purpose a large number of papers have been devoted (3, 9, 10, 11, 12, 13), and one of the central issues that has evolved deals with the problems of actuarial or statistical versus individual or clinical approaches to prediction. We will examine this problem as it relates not only to the Rorschach but also to a general concept of measurement theory and practice. It seems apparent that the use and extent of reliance upon various psychological measures is a by-product of how this problem is perceived and resolved by the individual social scientist; it also greatly influences the framework of the research that is carried on.

Probability appears to be of central importance in making predictions. If predictions are made on the basis of past events, wherein the conditions and frequencies under which an event has or has not occurred have been recorded, we are using a frequency concept of probability. Or predictions may be made that are based, not upon past enumerations of an event's specific occurrence or nonoccurrence, but rather on the basis of logical manipulations of factors considered relevant and our understanding of their relationships to the event being studied; this may be considered a logical



concept of probability. To illustrate these two concepts of probability, one may consider the problem of predicting the success or desirability of psychotherapy for a given patient. If it is known, for example, from past data, that patients with certain scores on given psychological tests have shown improvement through psychotherapy in eighty-three out of a hundred cases, one can use a frequency concept of probability in recommending psychotherapy for patients with such scores on psychological tests. When, on the other hand, we have a patient who we feel will not be able to tolerate stress, is of very limited intelligence, and the like, we might predict his chances for improvement through psychotherapy as "very poor" if we believe factors such as stress tolerance and intelligence to be related to the outcome of psychotherapy. This, then, would be a probability statement ("very poor") based upon logical relationships of factors.

There appear to be inherent limitations to both approaches. Individual behavior, which usually does not provide data on previous conditions and frequencies of occurrence, does not lend itself very readily to mathematical procedures of probability. Conversely, logical procedures of probability do not involve repetition of an event and therefore do not lend themselves to experimental verification, since verification cannot take place without repeated measurement.

Proponents of logical procedures of probability have argued that statistical prediction tends to distort because it leads to overgeneralization from limited data or to excessive narrowing of the application of specific data (2). They object to quantification because it must eventually be converted into linguistic forms and must be related and limited to the original context, which is attained through words and not numbers. To this purpose Lewin stated, "... the importance of a case, and its validity as proof, cannot be evaluated by the frequency of its occurrence . . ." (5, pp. 41-42).

Viewed in another light, it seems apparent that when a single case becomes a member of a class, we have, in essence, events that lend themselves to verification through a statistical approach to probability. That this is essential is attested to by the fact that clinical data not subjected to a statistical probability analysis do not improve the accuracy of prediction made from that data. The prediction can be verified and made more accurate only by ordering the event to a class and then enumerating the number of successes.

Although it is our contention that the specific approach to predicting behavior is quite significant in determining whether one relies on a statistical or clinical orientation, this is an artifact of the sophistication of the individual researcher and the available techniques rather than any inherent dif-

ference in the assessment of behavior. To examine this position further, a brief discussion of some of the major differences between so-called objective and projective tests should be useful.

#### TYPES OF TESTS

In an objective test, the item pool is so selected and determined as to limit to a very specific context the responses available to the subject. The examiner begins with a circumscribed and limited universe and determines whether the subject does or does not belong to this universe—and possibly the “extent” to which he belongs. Furthermore, because of the predetermined scoring system, the meaning of each item is fixed by the examiner rather than the subject. An implicit assumption, of course, is that the subject, the examiner, and the test constructor all have the same meaning for each item. The ambiguity value of the test item is, or should be, rigidly controlled and kept to a minimum.

Projective tests do not start with a well-defined universe of behavior and must rely on items whose meaning is determined by the subject. It is exactly this value that is sought and evaluated. Although projective tests are conventionally considered as unstructured, this is not altogether accurate. To begin, a wholly unstructured stimulus and/or situation, would not, almost by definition, be capable of eliciting any response whatever. Furthermore, when considering a test situation, we are dealing with the ambiguity (structure) of both the test situation as well as the stimulus. To describe a Rorschach examination as unstructured appears inaccurate, for the instructions given the subject and the expectations communicated to him are quite explicit. He is, for example, to respond verbally to the ink blots, there are numerous responses possible (and therefore expected), and he is encouraged to be spontaneous rather than selective.

A second distinction between objective and projective tests, which is almost a corollary to the previously mentioned one, is that the stimulus value of the material presented is intended to be limited in objective tests. The lack of ambiguity of test items and instructions are examples of the examiner's efforts to limit and control the number of cues confronting the subject. Again the stimulus value of each item has been predetermined as much as possible and is perceived as an extension of a defined universe. In projective tests the opposite effect is sought—namely, by providing a minimum of cues to the subject, to force him to seek his own experiences and perceptions for cues to utilize in responding. But it is misleading to think of the stimulus as completely ambiguous, thereby permitting the subject an infinite choice of responses. Although the choice is considerably greater than in an objective



test, the limitations of choice are attested to by our ability to categorize responses with considerable consistency.

A third major distinction between projective and objective tests is that the latter, by virtue of specific, predetermined objectives, have empirically determined units of measurement. It then becomes the function of the test to describe the individual not as an individual, but rather in terms of his approximation to the typical performance of culturally determined tasks for which an arbitrary, but internally consistent, empirical rating scheme has been devised.

It appears that the major limitation in the use of objective tests is their very limited application to a well-defined universe of behavior. They are therefore unable to evaluate the individual in terms other than the ordering of his behavior on the continuum of behavior predetermined by the selection and description of a particular universe. Paradoxically, it is this very limitation that made it possible for objective tests to achieve the degree of respect and usefulness that they now enjoy in many settings. By employing a basically reductionist approach to behavior, it has been possible to develop tests that survive scrutiny for validity and predictive ability. Projective tests are so labeled not only because of the greater ambiguity of their stimulus value, but also because they do not, and possibly cannot, proceed from a well-defined universe of behavior. They consequently are so devised as to maximize both their stimulus value and the variability of responses possible. It then becomes our task to order the obtained responses into the appropriate behavioral universe.

The general principle distinguishing these two approaches is found in the distinction between sign and sample tests (1). The majority of tests within psychology have been developed to sample an arbitrarily defined universe of behavior. The universe is the criterion by which we evaluate the test; the validity of the test is determined by the extent to which performance on the test is similar to the performance of the larger area of similar behavior from which the samples have been drawn. In this situation the criterion is regarded as the independent variable and the only question is: How well does the test correspond to the criterion by which it is to be judged?

A test may also be viewed as a sign, and one is then confronted with deciding to which universe the sign points. In this approach, we begin with observable behavior and try to understand it by discovering the broader context to which the signs belong.

The situation in these two types of test construction and test interpretation appears particularly relevant to the Rorschach technique where the crucial problem appears to be defining the particular universe to which the



signs point rather than selecting a sample from a predetermined universe. A further source of confusion directly relevant to the Rorschach is that, although the test is based upon signs that then have to be identified with a universe of behavior, when the Rorschach is scored, the theoretical position is changed to that of a sample test, with the assumption that we have a well-defined universe of behavior of which certain scoring categories are samples.

#### CHARACTERISTICS OF GOOD MEASURING INSTRUMENTS

The adequacy of any instrument of evaluation must be determined in terms of the purpose for which it has been constructed. Regardless of purpose, however, there are characteristics of measuring instruments that are desirable and applicable to all measurement situations. These characteristics then become the criteria with which to evaluate any given instrument. A good measuring instrument should, above all else, provide measures which are free of error. The types of error possible are described as compensating and biased ones (15). Compensating errors arise because an infinite number of measurements is not available. It is therefore possible to have errors that result from subject fluctuations at the various times measurements are obtained, errors made by a single rater or judge, and errors from various samples of behavior or test items. These types of errors have a tendency to cancel each other out when an infinite number of measurements is made by an infinite number of judges or when an instrument consists of an infinite number of items. Obviously, then, we are dealing with reliability errors.

Biased errors do not cancel out with repeated measurements. They are a consequence of the failure to reduce any given factor to a level of homogeneous characteristics or the failure to weigh these characteristics proportionally to their contribution to the general factor being measured. Whenever a measuring instrument does not measure the variable for which it is intended or does not contain an adequate cross section of items, biased errors are likely to result.

A good measuring instrument should be adaptable to the range of the variable studied. It is equally important that the units of measurement throughout the range being measured be equal. Although psychological measures usually do not achieve this objective, it is important to understand score differences at any point of the measuring instrument and to be able to determine whether, for example, differences at the lower end of a scale have the same meaning as equivalent differences at the upper end.

A further characteristic of a good measuring instrument is that it yields absolute scores. This means that the zero point of the instrument is equiva-

lent to the zero point of the factor being measured. Although the relative importance of this characteristic is often exaggerated, failure to satisfy this characteristic limits the usefulness of the instrument in inferring the location of the point of "no amount" of the factor being measured and also makes it impossible to interpret ratios between scores.

Some of the remaining characteristics of a good measuring instrument, which, although more "mechanical" in nature, are nevertheless essential, are: the instrument should be administratively feasible to construct and use; it should be available in duplicate forms which are equivalent; it should be able to measure changes at a level sufficiently sensitive for the factor being studied; and it should have adequate normative data. Although there are many aspects of normative data that are necessary and important, the comparability of the group being studied to the standardization group, as well as the size and adequacy of the latter group, are of prime consideration.

In summary, the adequacy of a measuring instrument may be evaluated in terms of the following types of errors:

A. Compensating errors resulting from:

1. Failure to obtain an infinite number of measurements.
2. Failure to obtain an infinite number of judges in scoring responses.
3. Failure to include an infinite number of test items.

B. Biased errors resulting from:

1. Failure to reduce the trait to homogeneous characteristics.
2. Failure to weight scores on items in proportion to their importance.
3. Failure to obtain an adequate cross section of items.
4. Subject dishonesty or incompetence.

## VALIDITY AND RELIABILITY

We shall define validity as the degree to which both compensating and biased errors are absent, while reliability we have described as the degree to which compensating errors alone are absent. A point to bear in mind is that these two concepts cannot be measured directly. Since they are based upon the absence of errors, it is assumed that true scores can be obtained; this is a necessary assumption, but not a true one. Were it otherwise, it would not be necessary to speak in terms of probability values. However, it is possible to determine estimates of both reliability and validity—to a considerable extent by the specific techniques employed in making our estimate. It is to this purpose that we shall now turn our attention.

Predictive validity refers to an instrument's ability to predict accurately behavior that is to be measured in the future. The usual technique for obtaining an estimate of this type of validity is correlation of test results with a



subsequent criterion measure; the extreme importance of obtaining an adequate criterion is apparent. The unreliability of the criterion as well as that of the measuring instrument will affect the latter's predictive effectiveness.

Concurrent validity is very similar to predictive validity, the major difference being that the criterion measure is obtained concurrently with the test performance. It is possible to have a high degree of concurrent validity but little predictive validity. This is true because of the nature of problems and criterion measures to which the two types of validity are applied. Estimates of both types of validity, by virtue of usually being expressed in the form of a correlation coefficient, fluctuate as a function of the homogeneity of the group with respect to the variable being studied. Homogeneous groups yield lower correlation coefficients than do more heterogeneous groups.

Construct validity differs from the other types of validity in that no definitive criterion measure is available. The emphasis is upon the theory that underlies the instrument. We first make predictions from the theory and then gather data with which to evaluate the predictions and, indirectly, the theory. The paucity of adequate criterion measures in clinical work has made construct validation of particular significance to psychologists working in this area.

Content validity refers to the adequacy with which the instrument samples content about which inferences are to be made. Although quantitative evidence is usually not readily available, the adequacy with which the instrument samples the population of behavior it is trying to measure can usually be taken as an appropriate indication of the instrument's content validity.

The above described types of validity are those which appear in the *Technical Recommendations* of the A.P.A. (14). They have been criticized on the basis that they are, with the exception of construct validity, excessively criterion-oriented and too specific in purpose (6). Gulliksen (4) summarized this position in stating that "The validity of a test is the correlation of the test with some criterion" and that any given test has many different validities, determined by the various uses of the test. Lord (8) is also a proponent of this position and rejects an over-all concept of validity, considering validity a specific quality. Loevinger (7) argued against such concepts of validity. She recognized only construct validity as scientifically useful. The construct she speaks of is equivalent to a statistic, serving as the best estimate of a parameter, or, in the case of construct validity, a trait. A measure of construct validity then becomes the estimate of behavior on the basis of test performance, taking into adequate consideration the inadequacy of predicting a single external criterion. The meaning of this notion of con-



struct validity is difficult to establish for projective tests. Considering projective test scores as intervening variables and single-item responses as an original datum, the difficulty becomes apparent if one considers structural validity as relationships among data and not intervening variables.

Reliability, as previously explained, may be defined in terms of the degree to which an instrument is free of compensating errors. These errors may occur because of fluctuations from repeated administrations of a test to a given group; the extent of these fluctuations or coefficients of stability is the estimate of reliability. One of the major considerations necessary in interpreting such a coefficient is that the obtained reliability, as a result of restricting the fluctuations of sampling test items, will tend to be an overestimate of the true reliability. Furthermore, the necessary assumption that repeated measurements will not be affected by virtue of the repetition is a tenuous one, which probably cannot be supported empirically.

A correlation between scores from two equivalent forms yields a coefficient of equivalence. Here again it is necessary to keep in mind that the obtained coefficient would have been lower had the compensating errors due to sampling test items been allowed to fluctuate as though truly different test items were provided. However, if the two forms are truly equivalent, the items will have been matched for difficulty and will increase the obtained measure of reliability. It has also been pointed out (6, 7) that since equivalent forms have, by definition, equal means, standard deviations, and mutual intercorrelations, their equivalency and reliability is estimated by the same set of correlations. The circularity of such a definition is apparent when we realize that the correlation or reliability we seek to estimate is used in defining two tests as equivalent.

The coefficient of internal consistency is the relationship based on the internal analysis of the instrument, usually dividing the test into two subtests, one composed of the odd items and the other of the even ones. The measure of reliability obtained from this method represents reliabilities for only half the test. It is then necessary to apply an appropriate formula (the Spearman-Brown formula) to predict the expected reliability of a test twice the length of either the odd or the even half-test.

The magnitude of the above described measures of reliability, like the size of all coefficients of correlation, is a function of the range of the characteristic being measured. If the instrument is applied to a group less homogeneous than the one for which it was devised and on whom it was standardized, the obtained measure of reliability will be considerably reduced. To interpret properly a measure of reliability obtained by any of these methods, it is necessary to consider the assumptions that (a) test items selected are sampled from an infinite population of possible items, (b) all

types of compensating errors are free to fluctuate in each form of the test, and (c) errors existing between either alternate forms, or odd and even items, are compensating and biased errors between forms of the subtest are entirely absent.

### *Clinical Applications of the Rorschach*

Since its introduction in 1921, the Rorschach has attained remarkable application and stature, enjoying now a firm position in the repertoire of techniques employed in the understanding and prediction of human behavior. As is often the case, clinical needs have taken precedence over experimental evaluation, and what began as a matter of expediency (use of this technique before it had been adequately assessed) has now become accepted practice. Indeed, the long history of the Rorschach's clinical use is an implicit deterrent for its further evaluation, for the recently trained clinician, better trained as a researcher and more disposed in attitude toward such an approach, is faced with the task of challenging years of use and accumulated "insight through experience."

Those familiar with the Rorschach literature are undoubtedly aware of some of the more frequent shortcomings of the research in the area. Samples that are woefully inadequate in size are used; samples are used to obtain data that have been applied to subjects from apparently different populations; statistical techniques are employed with considerable disregard for their theoretical and mathematical assumptions; constructs are postulated in such a manner that the transition from the theoretical position to the Rorschach data is tenuous, making subsequent generalizations unlikely to be supported by cross-validated studies. However, in spite of its limitations, the use of the Rorschach is gaining in frequency, and the judgments made on the basis of it appear, by and large, to be more respected and, at times, sophisticated. Unfortunately, they still remain unverified, most of the support coming from the clinical judgment of "experts."

Some investigators have sought internal validity for the Rorschach through attempts at finding psychometric groups corresponding to psychiatric ones, testing for criterion discriminations, evaluating the extent of judge's agreement, and inquiring into the basic assumptions underlying the Rorschach and its uses. Many of these assumptions are not explicitly stated in the literature and can only be inferred from traditional practices and interpretations.

External validation has also been sought in many ways, from testing Rorschach factors related to other tests, both of a projective as well as non-



projective nature, to seeking Rorschach factors related to behavioral manifestations of anxiety, suggestibility, impulsiveness, or diagnosis and prognosis. Comparisons have been made with such techniques as the Levy movement cards, Minnesota Multiphasic Personality Inventory, Bender-Gestalt, and Behn-Rorschach. The Rorschach has been tested for relationship with stress and emotional control, specific determinants such as color and social acceptance, and combinations of determinants with various personality traits.

A search of the recent literature indicated that, with the possible exceptions of child studies, relatively little effort is being devoted to developing adequate normative data for the Rorschach. Most of the studies in this area deal with special groups of subjects, such as preschool children or retarded adults, and/or with special problems, such as color, experience balance, and temporal development. The suggestion that Rorschach norms should describe adequately the samples upon which they are based, explicitly state whether the norms are based on a psychiatrically or statistically normal group, and be presented in terms of scoring categories has been almost totally ignored. Bearing in mind the methodological problems and shortcomings of much of the Rorschach literature, certain conclusions regarding scoring categories seem justifiable. The studies on color do not seem to support the general theory pertaining to the use of color by a subject. Contradictory results have been obtained in studies dealing with the depressant effects of color on reaction time. Lack of color in the usually colored cards was not found to depress the consistency of responses to the cards. Color shock has been present to a great extent in samples of normal subjects.

In the area of movement, extensor M responses have been said to reflect assertive personality types and to be found more frequently among male than female subjects. Those studies seeking to measure intelligence by means of the movement response have found an average correlation of approximately .40 between M responses and outside criteria of intelligence. The safest, and probably most accurate appraisal of studies in this area would indicate the absence of any one-to-one relationship between movement subcategories and behavioral variables.

In the studies concerning space, the findings are nebulous and generally not in support of the usually accepted theory of space percepts. Correlations between space responses and external criteria of oppositional tendencies range from approximately .25 to .45. Studies on shading have not dealt with the theoretical rationale for the scoring of shading.

The area of popular responses is characterized by considerably more theorizing than experimental verification. The raw-number method appears



to be equally valid as the percentage of popular responses in determining the adequacy level of a protocol.

Organizational factors do not appear to be related to the use of generalization on structured tasks, with correlation coefficients averaging approximately .35. Computing weighted Z scores does not appear to provide any advantage over noting undifferentiated Z responses.

The over-all effect of studies dealing with determinants is one of caution. This is also true of the efficacy of the Rorschach in differentiating psychiatric groups from each other and/or normal groups. Regarding schizophrenia, for example, there is no single set of scoring categories that consistently differentiates this group from others. Signs such as confabulation, contamination, high variability of form quality, and greater frequency of original responses are often associated with schizophrenics. Such signs as rare details, use of color, quality of form, and position responses, although also often associated with schizophrenics, are also found frequently with other psychiatric groups.

The major impressions regarding clinical use of the Rorschach are the following:

1. Used individually, it does not permit us to make very definitive statements about an individual. Even when used in conjunction with other clinical tests, Rorschach data do not yield unequivocal results, since meaningful interpretations depend more extensively on the examiner's skill and experience than upon the data itself.

2. It is based upon theory and assumptions not sufficiently evaluated. The apparent disregard for this truism appears to have resulted in a very considerable amount of very poor research which has been carried on with the apparent assumption of a well-evaluated and sufficiently validated theory behind it.

### *Outline of the Book*

In the preceding pages the reader should have been oriented toward the selections that will follow. It will be readily apparent that the Rorschach is not being examined within a specific theoretical or conceptual framework of personality. Indeed, it is our contention that to do so would be premature and severely limiting in the attempt to understand the perceptual processes that take place in the course of a Rorschach examination.

The book is divided into the following areas:

*Introduction:* My article has two major purposes. First, it attempts to make explicit some of the relationships that exist between the field of tests

and measurement and the process of personality measurement, emphasizing such basic distinctions as statistical and clinical approaches to the prediction of behavior. Second, it summarizes very briefly the results of Rorschach research that has considered clinical problems.

Mooney concerns himself with the relationships that exist between the field of perception and the assessment of personality. He considers some of the basic relationships that are postulated between the two fields and makes it possible for the reader to make inferences about the areas needing research.

*The Projective Method:* The article by Frank may well be considered the fundamental exposition of the process of projection and its importance in assessing test results and behavior. Although our understanding of these phenomena has expanded considerably since the writing of this article, the foundation it provided appears to have withstood the critical evaluation of many years of research and thought.

The article by Sargent presents a comprehensive and relatively detailed treatment of numerous projective techniques. For each technique discussed she presents a summary of the theoretical background and appropriate experimental data. In addition, she provides the reader with an extensive bibliography, divided according to the specific projective methods to which it applies. This article is intended to give the reader a historical perspective of the need and subsequent development of the Rorschach.

The article by Lord supports with experimental data the importance of such factors as set and the subject's perception of the test situation upon the protocols obtained. Although numerous studies have dealt with this problem, this article appears to be particularly useful because of its excellent methodological approach.

*Scoring:* Whether or not one should score a Rorschach protocol remains an unresolved question, with solutions ranging from clinicians who offer a categorical "no" to those who adhere to the most elaborate and detailed scoring schemes. The first study by Wittenborn fails to support the frequently made assumption of the behavioral similarity of responses falling within a given Rorschach scoring category. Furthermore, the assumption of psychologically significant differences between responses assigned into different scoring categories is also not supported. The second article by Wittenborn offers evidence for the need to re-evaluate the use of certain major scoring categories and the emphasis often attached to some of them. Wittenborn's third article, in addition to yielding experimental data upon the meaning and interrelationships of the major Rorschach scoring categories, also discusses some possible approaches to the validation of the Rorschach. Of more practical significance are the results obtained by Knopf, which in-



dicate the futility of utilizing Rorschach summary scores in differentiating psychiatric groups.

*Validity:* This concept remains as a crucial issue in the use of the Rorschach for purposes other than research. The paper by Schneider outlines some of the methodological and theoretical considerations that might be used in making the Rorschach a more valid measure. Specifically, he views the problem as one of "relating Rorschach variables to independent measures of component personality measures."

The article by Palmer presents two methods for Rorschach validation. He concludes that neither of these is completely adequate for the task, but, used in conjunction with a design suggested by Cronbach, they may make it possible to obtain a satisfactory method for the validation of projective techniques.

Cronbach, reviewing much of the Rorschach literature that has sought statistical evaluation of Rorschach theory, concludes that methodological errors have led to errors of omission in establishing significant relationships between test and behavior variables.

*Reliability:* The relative paucity of research in this important area is difficult to understand. The study by Hertz, although of considerable vintage, still remains as one of the better sources of verification for the reliability of the Rorschach. The problem of scorer reliability appears to have been even more overlooked. Dana's study, although basically compatible with the assumption that it is possible to have scorer reliability with the Rorschach, yields data indicating that individualized scoring systems and ambiguous scoring categories tend to reduce scorer reliability considerably. The last study in this section provides experimental data to support the cautions recommended in considering a single Rorschach examination as an adequate or stable personality description.

*Current Status:* In this section an attempt is made to estimate the position of the Rorschach in the field of psychology. The article by Thurstone appears to be an indictment of the Rorschach on the basis of its wide applications without sufficient recourse to experimental evidence to justify such use. Zubin is more specific in his over-all evaluation, concluding that the Rorschach is an interview whose content is of potential usefulness in evaluating behavior. However, the presently employed techniques for analyzing this content are inadequate.

The article by Cronbach and Meehl appears to be particularly relevant to the development of valid procedures in the use of the Rorschach. It offers an extensive discussion of construct validity, which appears uniquely appropriate to the Rorschach.

*Summary:* The article by Hertz reviews a considerable number of studies



concerned with the Rorschach. Her evaluation of these is somewhat more optimistic than that of others who have considered the same body of literature. Although she recognizes the need for more research data, replication of important results obtained through seemingly adequate research designs, and more appropriate statistical techniques, she feels that many important Rorschach hypotheses have been verified and the "test works," failing only when "subjected to piece-meal and rigid statistical manipulations."

## REFERENCES

1. ANASTASI, A. *Psychological testing*. New York: Macmillan, 1954.
2. BROWER, D. The problem of quantification in psychological science. *Psychol. Rev.*, 1949, 56, 325-333.
3. COTTRELL, L. S. The case study method in prediction. *Sociometry*, 1941, 4, 358-370.
4. GULLIKSEN, H. *Theory of mental tests*. New York: Wiley, 1950.
5. LEWIN, K. *A dynamic theory of personality*. New York: McGraw-Hill, 1935.
6. LOEVINGER, J. Objective tests as instruments of psychological theory. *Psychological Reports, Monograph Supplement* 9, 1957.
7. LOEVINGER, J. The attenuation paradox in test theory. *Psychol. Bull.*, 1954, 51, 493-504.
8. LORD, F. M. Some perspectives on "The alternation paradox in test theory." *Psychol. Bull.*, 1955, 52, 505-510.
9. LUNBERG, B. A. Case studies vs. statistical methods—an issue based on misunderstanding. *Sociometry*, 1941, 4, 379-383.
10. MEEHL, P. *Statistical vs. clinical prediction*. Minneapolis: University of Minnesota Press, 1954.
11. QUEEN, S. A. Social prediction-development and problems. *Sociometry*, 1941, 4, 371-373.
12. SARBIN, T. R. The logic of prediction in psychology *Psychol. Rev.*, 1944, 51, 210-228.
13. STOFFER, S. A. Notes on the case study and the unique cases. *Sociometry*, 1941, 4, 319-357.
14. Technical recommendations for psychological tests and diagnostic techniques. *Psychol. Bull. Suppl.*, 1954, 5, No. 2, Part 2, 1-38.
15. WERT, J. E., NEIDT, C. O. AND AHMANN, J. S. *Statistical methods in educational and psychological research*. New York: Appleton-Century-Crofts, 1954.

*Bernard Mooney*

## PERSONALITY ASSESSMENT AND PERCEPTION

**T**HE PROVING grounds for testing out the innumerable hypotheses regarding personality assessment and perception have typically been identified with the theory and uses of the projective technique. Judging simply from the voluminous body of research that has grown up around the Rorschach test, it has become the prime vehicle for assaying the verifiability of the many proposed functional linkages between perceptual processes and personality structure.

In the clinical setting, the projective techniques are generally employed to provide a dynamic view of the individual—that is, to afford a hierarchical ranking of the individual's needs (intensity) as well as, perhaps, indications of the objects and events by which these needs are initiated and terminated (direction). Again, in the clinical setting this dynamic view, so obtained, is employed as a data pool that, when integrated, supplies information about the subject not available by other testing techniques.

Undaunted by faint memories of negative research findings regarding the uses of projective techniques, the practicing clinician forges ahead, collecting score upon score of Rorschach protocols. With such data at hand, he will go on to formulate answers to: "Will this patient attempt suicide?" or "Will this person decompensate under present pressures?" Interestingly enough, in the many controversies over the uses of projective techniques, the practicing



clinician has been the proverbial straw man, abused as though he and he alone were entirely responsible for the development of these tests. At the least, he is accused of perpetuating their existence. The clinician is not an opinionated, ill-informed individual who turns his back on the facts of the case. Rather, his continued use of such techniques reflects an awareness of an incontrovertible, though perhaps nonverifiable, fact that "These tests can tell me much about the individual's personality."

If the clinician were to make explicit the premises underlying the utilization of the Rorschach test and express them as a syllogism, it would assume a form such as: The Rorschach test is a perceptual task; personality traits are known to determine, in part, perceptual processes; therefore, from a consideration of the operation of perceptual processes as revealed in the Rorschach, the clinician can obtain useful information about an individual's personality structure.

Perhaps there may be objections to the choice of terms employed—"traits," "personality structure," and so on—as hypothetical constructs with much surplus meaning. The syllogism is not proposed as a prime example of the hypothetico-deductive method; it is simply an attempt to represent the implicit understanding most clinicians share in their acceptance of the Rorschach test as a useful assessment technique. Furthermore, the approaches to Rorschach research appear to share, in varying degrees, this outlook toward the test in regarding it as a tool for uncovering new facts about personality dynamics and behavior.

This paper will be devoted to a critical scrutiny of the elements that make up this rationale, with the hope that some insight may be gained into the problems that arise in attempting to establish the Rorschach projective techniques as empirically sound assessment procedures.

In the present paper the Rorschach test will be referred to as an "assessment technique," as distinguished from a "psychometric technique," following the distinction introduced by Cronbach (1960). The psychometric technique insists upon the reduction of test behavior to quantified score units that are combined statistically in predicting a specific criterion. On the other hand, assessment procedures emphasize reliance upon the logical evaluation of test observations combined in a quasi-artistic manner by the clinician. Conceived in this fashion, the problem of assessment techniques does not lie entirely in the unique characteristics of the tests employed (what data are employed), but rather includes the procedures utilized in the clinical application of assessment techniques (how the data are employed). It is not simply a matter of developing and perfecting the tests themselves but necessarily involves the integrative process of the clinician—that is, the area of clinical judgment. It will be the purpose of this article first to con-

sider the import of the premises outlined above in regard to the Rorschach itself, and second, to suggest how these considerations are related to the activity of the clinician.

The major premise indicates that the Rorschach is a perceptual task. For purposes of the present discussion, "perception" will be defined simply as the behavioral process which falls between dependence upon stimuli (sensation) and independence of stimuli (conceptualization or thinking) whereby the individual represents reality to himself. The range of behavior patterns that are included under the label "perception" are not considered to cluster around some point midway between pure sensation and cognitive activity, but rather to extend to both extremes of this imaginary continuum.

Perceptual behavior involves the characteristics of the perceiver as well as the characteristics of the stimuli, so that dependence upon or independence of the stimuli can be influenced in terms of the condition of the perceiver. Now, while one may insist, along with Rorschach (1942, 1951), that the response to the ink blots is a perceptual process, it is evident that the test situation is markedly different from classical perceptual tasks, such as the Gestalt Phiphenomenon or figural aftereffect. It is the distinguishing characteristic of the projective technique that the value of external stimulus cues is markedly reduced so that, ostensibly, the subject is forced to rely predominantly upon the promptings of his own internal cues. According to the projective hypothesis, these internal cues consist basically of the individual's personal needs, drives, and motives. It is further assumed that it is these internal cues that, in some measure, determine perception in the real life situation. More important, it is assumed that these internal cues may be utilized in explaining or predicting the subject's overt behavior, that the organism responds to the external environment according to his perception of it.

In the projective test situation, then, the ambiguity of the stimuli as well as that of the set-determining instructions foster reliance upon cognitive processes. In the classical perceptual task, however, the entire stimulus complex demands reliance upon sensory processes.

Another way of stating the distinction is that, in the classical perceptual task, the subject is required to make some act of discrimination; that is, his attention is focused upon the sensory elements in the perceptual process. In the Rorschach, however, the subject is required to minimize the sensory elements of the perceptual process and to emphasize, rather, the imaginal or associative elements; he is encouraged to depart from the concrete characteristics of the blot. Were this not so, the subjects' responses could hardly be other than "ink blots," or "smears," or "blotches of paint." To what extent, then, do the facts of perception described in the classical experiment



apply to Rorschach behavior? It is not sufficient to indicate merely that they apply with some modifications.

Consider for a moment the typical perceptual variables employed in an analysis of test performance, i.e., the Rorschach scoring categories. These were originally developed by Rorschach himself, who offered no clear rationale for their creation other than their apparent utility in distinguishing among the test records of the various diagnostic groups; they have been little altered down to the present. One becomes involved in a circular validation process unless the existence of such variables can be demonstrated in tasks other than the Rorschach itself. At present, there is no clear-cut linkage between perceptual task variables and the Rorschach categories. In fact, it appears that an empirical basis for the distinctness of the categories is lacking (Murstein, 1960).

Accepting for the moment that scoring categories such as F, W, and M may be related to perceptual variables in some fashion, how is one to relate these to the established principles of perception? If one accepts the standard interpretation of the W response as indicative of organizational activity, then he might predict that W responses would seldom occur on blots that lack the elements of proximity, similarity, and the like, conducive to the perceived cohesiveness of parts. Or, if one accepts Orlansky's findings (1942) that angulated figures lend themselves readily to Phi-phenomena (are readily perceived in motion), he might predict that movement responses would be elicited more readily on blots exhibiting angularity.

Another example might be the phenomenon of atmospheric perspective. Since blurred, indistinct objects are perceived as more distant than clearly outlined ones, blots exhibiting both vague and clear outlines might be predicted to give rise to vista responses. Further, since no blot represents only one of these stimulus features, predictions might be formulated regarding the hierarchy of influence of these factors in determining responses. It is precisely this type of research, attentive to the stimulus value of the blots, that appears most important.

The above are simply indications that it is possible to relate the known facts of perception to performance upon the Rorschach. As is evident, Rorschach himself derived the response categories apart from a consideration of the stimulus value of the blots and of the perceptual factors involved. The importance of filling this void is apparent when one realizes that apart from the test itself these scoring categories have no meaning. With the empirical validation of such relationships between scoring categories and perceptual principles, the voluminous body of Rorschach research may more easily be incorporated into psychological theory. This appears to be the most basic step in a profitable approach to the utilization of the Rorschach. Without



Date... 14-3-85

Acc. No. 3051

such empirical underpinnings, the entire structure of projective tests of this type remains unstable indeed.

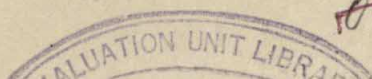
On the level of the integrative activity of the clinician, he is forced to classify the subject's responses according to the quasi-perceptual scoring categories. The subject who has not been directed to concentrate upon the specific determinants of his response is usually not capable of delivering the precise introspective report which the clinician desires. While the clinician ostensibly analyzes the subject's perceptual processes into their components, identification of this or that determinant depends largely on the vagaries of the subject's word choice. At this point, one might also pose the question about the extent to which the clinician's personal perception of the blot may contaminate the selection of scoring determinants in the typical response. In other words, may not the clinician learn to assume that in a given population, most animal responses to Card II are large-detail, form-determined responses? In the absence of clear verbalization to the contrary, any animal responses may readily, though perhaps illegitimately, be scored as DF. Finally, the meaning of the scoring categories may be further contaminated by the fairly common practice of assuming that, in the absence of a specific reference to other determinants, a response may be considered to be form-determined.

In summary, the major problem involved at the level of clinician activity in the utilization of scoring categories arises from the fact that such categories depend directly upon the clinician's interpretation of the verbal report of an untrained introspectionist. The ambiguity inherent in such cues fosters the reliance by the clinician upon internal cues such as previous experience with the test.

Regarding the minor premise of the syllogism—that personality traits determine, in part, the perceptual process—there can be little discussion. The precise meaning of this premise in terms of the Rorschach test, however, is another matter entirely. The statement would generally be taken to include all the stock-in-trade hypotheses that the practicing clinician accepts along with the test itself. Interestingly, these hypotheses have changed little since their original presentation by Rorschach. They include: C responses are indicative of emotional instability; and the better the form visualization (F plus %), the better the emotional control. Rorschach derived these hypotheses from the predominance of the various response categories among his psychiatric and normal normative groups. For example, the subjects who gave the most C responses were the "epileptics, manics . . . or notoriously hot headed normals. From this it was concluded that C answers have a 'symptom value' . . . the tendency to impulsive emotional discharge" (p. 33, 1942, 1951). This is the tenor of the derivation of the Rorschach hypotheses.

137'842  
HIR

10033





Needless to say, continuing attempts to prove or disprove these hypotheses do not appear to be the most profitable approach to the use of this test.

Perhaps a more important consideration is the applicability of the stated relationship between personality traits and perception. Many of the studies that established the fact that personal needs, attitudes, and the like influence perception utilized tasks that occupied a position on the continuum between sensation and cognition, lying closer to dependence upon the stimuli; that is, either the stimuli employed were nonambiguous or the subjects were given a definite task to perform which oriented them directly to stimulus characteristics. On the Rorschach test, however, the subject is faced with relatively ambiguous stimuli and, certainly, with an ill-defined task that is generally assumed to be a "test of imagination" (8, p. 61). If this is true, can one legitimately accept the notion that needs and so on influence both types of perceptual activity in the same manner? Does it not appear reasonable to believe that needs influence the behavior of a stimulus-oriented subject differently from the behavior of the subject with an interpretative set?

Again, at the level of clinical judgment, problems are multiplied because of the lack of clarity in the meaning of perceptual processes as applied to Rorschach performance. In reality, the clinician's pool consists of verbalizations that describe the percept achieved rather than the process whereby it is achieved. Thus, while important contributions have been made toward an understanding of the relationship between abnormalities of the sensory processes in perception and constellations of the personality traits (Granger, 1960), the applicability of such research to a task like the Rorschach constitutes a rather knotty problem. To understand, for example, that the severely anxious subject may be less sensitive to the nuances of color when he is forced actively to discriminate colors is not to suggest that the subject who fails to verbalize color when describing his percept is severely anxious. Again, the uncontrolled nature of the response to the Rorschach task is quite at variance with the performance of a subject in a straightforward discriminatory task.

The final premise to be considered is that of the relationship between the Rorschach task and personality structure. The point of departure is the acceptance of a theory of dynamic interrelationships among quantifiable factors that determine overt behavior. These factors generally demonstrate the properties of hypothetical constructs rather than of intervening variables. In practical terms, the intervening variable is created by the facts; it is called into being to "name" an empirically verified relationship between antecedent and consequent conditions. The hypothetical construct, on the

other hand, is created within a logical, theoretical framework and may have no existence apart from the theory that called it into being.

In theory construction, of course, the hypothetical construct represents the early attempts to uncover the laws relating the data investigated. At this stage, the constructs possess marked heuristic value in pointing the way toward meaningful experimentation. In time, with the accumulated discovery of a body of lawful relationships, they may be confirmed as intervening variables. At the present stage of development in personality theory, most of the explanatory concepts remain at the level of hypothetical constructs. The use of such constructs reduces the efficiency of the predictive process, because the construct is not tightly bound to the consequents or overt behavior; nor is it specifically linked to the antecedents or test measures of the antecedents.

Thus, at the level of the clinician's attempts to predict behavior, the ambiguity of the personality constructs that he selects will markedly reduce his over-all predictive efficiency. In addition, his reliance upon the use of logically derived scoring categories automatically results in the loss of the highly valued "personal information" about the subject that the projective technique is constructed to yield.

It is proposed that this unique, personal information about the subject rests in the percept achieved—that is, in the content of the subject's verbalizations. While much lip service has been paid to the importance of content analysis, there has been a singular absence of controlled studies focusing upon the more cognitive aspects of perception in the projective techniques. That the content of the response is, at times, a more appropriate source of cues for predicting to a criterion than any system of logical categories is an implicit postulate underlying the cognitive activity of the clinician. The illustration of this fact cited by Hammond (1955) in the context of Rorschach protocols is indeed impressive. Given the task of estimating measured intelligence on the basis of scoring categories alone as opposed to the use of verbatim responses, the correlation between clinician estimates and measured intelligence rose from .47 to .64. As a matter of fact, the estimates by clinically inexperienced college sophomores based on verbatim responses correlated .58 with measured intelligence. It is not contended that a study of this type supports the validity of content analysis, since one may ask, "What is it really in the verbatim response that serves as a cue?" In addition, the study points out the fact that errors in prediction by clinicians arose partially from their erroneous weighting of scoring categories. This last point importantly relates to the objections raised by Holt (1958) in regard to the controversy over clinical judgment vs. statistical methodology. More will be said of these points later.



If the projective techniques are to be considered seriously as measures of perceptual processes, then an analysis of a technique that omits the area of meaning is incomplete. The use of logically derived score categories (as in the Rorschach) often increases the error variance in prediction in several aspects. Though the categories may be logically meaningful, the criteria for inclusion within a category may depend largely upon the value judgments or perceptions of the clinician himself. Hence, such systems prove statistically unreliable. Furthermore, granted a nonambiguous scoring system, the extent to which the categories represent empirically meaningful dimensions in the perceptual process is vitally related to the validity of clinical inferences regarding personality structure. Finally, predictions of behavioral criteria from such categories are mediated by their relationship to the personality constructs. The comments made earlier concerning the function of such hypothetical constructs are applicable here.

But why consider the dimension of meaning in projective testing? First, it is clear that any definition of "perception" necessarily involves, in some fashion, the integration of stimuli into meaningful configurations—that is, "perception" is the process through which one arrives at meaning. It would appear that "meaning" is an appropriate focus in any attempt to relate perception and personality. It is this realization that has stimulated research such as that by Elizur (1949) in rating content according to the degree of anxiety and hostility exhibited. Zubin (11) has concluded that content analysis generates accurate descriptions of personality dynamics. In addition, he (1955) has also demonstrated the utility of a scoring system based upon content analysis in that raters, utilizing the system, evidence a high degree of agreement. However, such approaches still remain one step removed from the area of subject meaning. With scales of the type devised by Zubin, the context of meaning is derived from the experimenter's frame of reference. This frame of reference may or may not be in accord with that of the individual responding. Such attempts to quantify the dimension of meaning are vulnerable to the same objections raised against present systems of categorizing responses.

In the present article, the logically derived scoring systems employed have been subjected to scrutiny, and their tenuous relationship to dimensions of perception has been considered. However, the most important problem in regard to these assessment techniques—that of the stimulus value of the ambiguous stimuli composing the various projective techniques—has not yet been treated. In the context of the logically derived scoring systems in present use, it has been important to consider how the various component elements of the stimulus are related to response tendencies. While research involving these stimulus elements has yielded vital information about the

effects of stimulus properties upon subject reactions, it has not touched upon the basic question of the perceptual meaning of the stimuli and their relationship to content. The former type of research is quite atomistic and at times seems in direct contradiction to the holistic principles underlying the personality-assessment techniques. The latter type of research has been lacking, though the techniques for attacking the problem of meaning do exist. The attempts that have been made with the Rorschach—for example, Zax and Laiselle (1960)—have not received the attention due them, perhaps because the results have been interpreted as either supporting or disproving the stock-in-trade clinical hypotheses about the meaning of the blots.

While the problems of the projective tests as methods of personality assessment are many and serious, they do not justify a negative attitude toward the techniques themselves. The function of personality assessment through perceptual processes is not theoretically unsound nor experimentally impossible. What is required, however, is a careful consideration of the problems associated with personality constructs, the definition of the perceptual process, and the methods of quantifying test performance. Basic to all of these points is the matter of stimulus value or the perceptual meaning of the projective stimuli. Without some empirical measures of the meaning dimension, the clinician is faced with an equation consisting of two unknowns, the task and the subject. It is precisely the second member of the equation which the clinician purports to define.

As was suggested earlier, there appears to be overlap between the problems of personality assessment by means of the projective techniques and the contribution of the clinician's judgment to assessment. It is important to understand clearly the unique problems introduced into assessment procedures by the clinician's reliance upon the more cognitive aspects of his own perceptual activity. There is perhaps no better way in which to summarize these possible sources of error than by referring the reader to Holt's (1958) description of predictive process. In his analysis, the problem areas in personality assessment and prediction of behavior through perceptual processes are clearly delineated.

In paraphrase of Holt's analysis, the first two steps require an understanding of the behavior to be predicted and, conjointly, the isolation of the intervening variables related to the criterion behavior. In the present context, these steps demand basic research resembling job analysis, discovering rather than assuming the environmental and organismic variables which interact in determining behavioral outcomes. While the behaviors that the clinician is asked to predict often represent "divergent phenomena" (Cronbach, 1960) and are not amenable to the predictive process, the vast majority



of his clinical decisions involve behaviors that represent "convergent phenomena" and that can be subjected to analysis.

In regard to the isolation of relevant intervening variables, the research task is more formidable. As was indicated earlier, the conceptual frameworks within which assessment techniques are interpreted are derived from the various personality theories replete with hypothetical constructs rather than intervening variables. This does not, however, render the research task impossible, though it does caution against unwarranted claims for relative superiority among such constructs. Their functions are basically heuristic, supplying suggestions for fruitful areas of investigation that may or may not eventuate in the development of personality constructs of the intervening variable type.

The third step involves the development of adequate measures of the important variables. In the context of personality assessment through perception, it is proposed that the dimension of meaning be a focus for research. Experimental studies have pointed up the utility of content or meaning analysis, though until recently no systematic procedures for the study of meaning have been utilized. Without some normative data regarding stimulus meaning, the clinician is faced with the dilemma of attempting to explain one unknown (the subject) by means of another (the stimulus).

The final steps in the predictive process are relating the measures to the intervening variables and actually combining the data to yield predictions. These steps constitute the test of the entire process. The results obtained at these levels serve as the standards for evaluating the practical validity of the entire experimental schema.

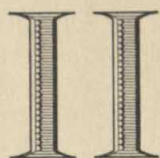
While Holt's analysis originally applied to the problem of statistical vs. clinical procedures, it also refers pointedly to the problem of projective techniques in personality assessment. The essential focus of his analysis, as well as of the present analysis, is the reasonable demand for the reduction of intuitive linkages between test results and overt behavior (reduction of the clinician's reliance on his own personal internal cues), and the search for lawful functional relationships firmly rooted in sound experimental techniques.

## REFERENCES

1. CRONBACH, LEE J. *Essentials of psychological testing*. (2nd ed.) New York: Harper & Brothers, 1960.
2. ELIZUR, A. Content analysis of the Rorschach with regard to anxiety and hostility. *Rorschach Res. Exch.*, 1949, 13, 247-284.

3. GRANGER, G. W. Abnormalities of sensory perception in Eysenck, H. (ed.) *Handbook of abnormal psychology*. New York: Basic Books, 1961.
4. HAMMOND, K. Probabilistic functionalism and the clinical method. *Psychol. Rev.*, 1955, 62, 255-262.
5. HOLT, ROBERT. Clinical and statistical prediction: a reformulation and some new data. *J. Abnorm. Soc. Psychol.*, 1958, 56, 1-12.
6. MURSTEIN, B. I. Factor analyses of the Rorschach. *J. Consult. Psychol.*, 1960, 24, 262-275.
7. ORLANSKY, JESSE. The effect of similarity and difference in form on apparent visual movement. *Arch. Ps. N. Y.*, 1940, no. 246.
8. RORSCHACH, HERMANN. *Psychodiagnostics*. New York: Grune & Stratton, 1951.
9. ZAX, M. AND LOISELLE, R. H. Stimulus value of Rorschach Inkblots as measured by the semantic differential. *J. Clin. Psychol.*, 1960, 16, 160-163.
10. ZUBIN, J. Failures of the Rorschach technique. *J. Proj. Tech.*, 1954, 18, 303-315.
11. ZUBIN, J., ERON, L. D. AND SULTAN, F. A. Psychometric evaluation of the Rorschach experiment. *Amer. J. Orthopsychiat.*, 1956, 26, 773-782.

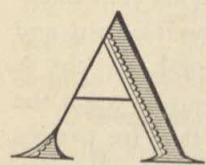




*The Projective Method*

Lawrence K. Frank

# PROJECTIVE METHODS FOR THE STUDY OF PERSONALITY



AN INITIAL difficulty in the study of personality is the lack of any clear-cut, adequate conception of what is to be studied. The recent volumes by Allport and by Stagner, and the monograph by Burks and Jones,<sup>1</sup> may be cited as indicators of the confusion in this field, where, as they show, there are so many conflicting ideas and concepts, each used to justify a wide variety of methods, none of which are wholly adequate.

A situation of this kind evokes different responses from each person according to his professional predilections and allegiances. Obviously pronouncements will be resisted, if not derided, while polemics and apologetics will only increase the confusion. The question may be raised whether any light upon this situation can be obtained by examining the *process* of personality development for leads to more fruitful conceptions and more satisfactory methods and procedures.

Reprinted from *J. Psychol.*, 1939, 8, 389-413, by permission of The Journal Press.

1. Cf. Allport, G. W. *Personality: A Psychological Interpretation*. New York: Holt, 1937. Cf. Stagner, R. *Psychology of Personality*. New York: McGraw-Hill, 1937. Cf. Burks, Barbara S., & Jones, Mary C. Personality development in childhood: A Survey of problems, methods and experimental findings. *Monog. Soc. Res. Child Devel.*, 1936, 1, 1-205.



Specifically, it is suggested that we reflect upon the emergence of personality as an outcome of the interaction of cultural agents and the individual child. In the space here available only a brief summary statement is permissible of the major aspects of this process in which we may distinguish an individual organism, with an organic inheritance, slowly growing, developing, and maturing under the tutelage of parents and teachers intent upon patterning him to the culturally prescribed and socially sanctioned modes of action, speech, and belief.

As elsewhere stated,<sup>2</sup> the child is not passive clay but a reacting organism with feelings, as are the parents, nurses, and teachers who are rearing him. He therefore receives training in the prescribed cultural and social norms of action, speech, and belief, according to their personal bias and feelings, and he accepts this training with varying degrees of observance, always idiomatically and with feelings toward these instructors. What we can observe then is the dual process of *socialization*, involving sufficient conformity in outer conduct to permit participation in the common social world, and of *individuation*, involving the progressive establishment of a private world of highly idiosyncratic meanings, significances, and feelings that are more real and compelling than the cultural and physical world.

The foregoing does not imply any subjective duality or other traditional dichotomy; it is an attempt at a simple statement of the well-known and generally accepted view that in all events we may observe both similarities or uniformities and also individual deviations. We may concentrate upon the larger uniformities and ignore the individual components that are participating, as we do in measuring the temperature, pressure, and other properties of a gas or we may look beyond the aggregate uniformities to the individual, discrete molecules and atoms and electrons which, as we now are realizing, are highly erratic, unpredictable, and far from that uniformity of behavior described statistically. Thus, we may observe a similar antithesis between the group uniformities of economic, political, and social affairs and the peculiar personal conduct of each of the citizens who collectively exhibit those uniformities and conformities.

Culture provides the socially sanctioned patterns of action, speech, and belief that make group life what we observe, but each individual in that group is a personality who observes those social requirements and uses those

2. Cf. Frank, L. K. Fundamental needs of the child. *Men. Hyg.*, 1938, 22, 353-379. Cf. Frank, L. K. Cultural coercion and individual distortion. *Psychiatry*, 1939, 2, 11-27.

patterns idiomatically, with a peculiar personal inflection, accent, emphasis, and intention.<sup>3</sup> Strictly speaking, there are only these individuals, deviating from and distorting the culture; but with our traditional preoccupation with uniformities we have preferred to emphasize the uniformity of statistical aggregates of all activities as the real, and to treat the individual deviation as a sort of unavoidable but embarrassing failure of nature to live up to our expectations. These deviations must be recognized, but only as minor blemishes on and impediments to the scientific truths we seek!

Those ideas flourished in all scientific work up to about 1900 or 1905 when X-rays, quantum physics, relativity, and other new insights were developed that made these earlier ideas obsolete, except in a number of disciplines which still cling to the nineteenth century. Thus it is scientifically respectable, in some circles, to recognize that uniformity is a statistical group concept that overlays an exceedingly disorderly, discontinuous array of individual, discrete events that just won't obey the scientists' laws! It is also respectable to speak of organization and processes "within the atom," although it is recognized that no direct measurements or even observations can be made within the atom; inferences being drawn from activities and energy transformations that are observable and frequently measurable.

For purposes of discussion it is convenient to see individuals (*a*) as organisms existing in the common public world of nature, (*b*) as members of their group, carrying on their life careers, in the social world of culturally prescribed patterns and practices, but living, (*c*) as personalities in these *private worlds* which they have developed under the impact of experience. It is important to recognize these three aspects of human behavior and living because of their implications for scientific study.

As organisms reacting to the environmental impacts, overtly and physiologically, human activity presents a problem of observation and measurement similar to that of all other organisms and events. The human body moves or falls through geographical space, captures, stores, and releases energy, and so on. As members of the group, individuals exhibit certain patterns of action, speech, and belief that may be aggregated into larger categories of uniformity or cultural and group norms; at least we find certain pronounced, often all-inclusive modes in their observed activities in which they tend to conform to social and cultural prescriptions.

When we examine the personality process or *private worlds* of individuals we face a somewhat peculiar problem, because we are seeking not the cultural and social norms of the uniformities of organic activity, but rather

3. Cf. Benedict, Ruth. *Patterns of Culture*. Boston: Houghton Mifflin, 1934. Cf. Mead, Margaret. *Sex and Temperament*. New York: Morrow, 1935. Cf. Bateson, G. *Naven*. Cambridge: Cambridge Univ. Press, 1936.



the revelation of just that peculiar, individual way of organizing experience and of feeling which personality implies.

In this context we may emphasize then that personality is approachable as a *process* or operation of an individual who organizes experience and reacts affectively to situations. This process is dynamic in the sense that the individual personality imposes upon the common public world of events (what we call nature), his meanings and significances, his organization and patterns, and he invests the situations thus structured with an affective meaning to which he responds idiomatically. This dynamic organizing process will of necessity express the cultural training he has experienced so that until he withdraws from social life, as in the psychoses, he will utilize the group-sanctioned patterns of action, speech, and belief, but as he individually has learned to use them and as he feels toward the situations and people to whom he reacts.

If it were not liable to gross misunderstanding, the personality process might be regarded as a sort of rubber stamp which the individual imposes upon every situation by which he gives it the configuration that he, as an individual, requires; in so doing he necessarily ignores or subordinates many aspects of the situation that for him are irrelevant and meaningless and selectively reacts to those aspects that are personally significant. In other words, the personality process may be viewed as a highly individualized practice of the general operation of all organisms that selectively respond to a figure on a ground,<sup>4</sup> by reacting to the configurations in an environmental context that is relevant to their life careers.

It is interesting to see how the students of personality have attempted to meet the problem of individuality with methods and procedures designed for study of uniformities and norms that ignore or subordinate individuality, treating it as a troublesome deviation which derogates from the real, the superior, and only important central tendency, mode, average, etc. This is not the occasion to review these methods and the writer is not competent to assess them critically, but it is appropriate to point out some aspects of the present methodological difficulty we face in the accepted quantitative procedures.

Since individuals, as indicated earlier, learn to conform to the socially sanctioned patterns of action, speech, and belief (with individual bias and flavor of their own), it is possible to establish the social norms appropriate for *groups* of like chronological age, sex, and so on and to construct standardized tests and to calculate statistically their validity, i.e., do they measure or rate what they are expected to measure or rate for each group; and their

4. Cf. Frank, L. K. The problem of learning. *Psychol. Rev.*, 1926, 33, 329-351.

reliability, i.e., how well or reliably do they measure or rate the performance of the groups?<sup>5</sup>

While standardized tests are generally considered to be measurers of individual differences, it would be more appropriate to say that they are ratings of the degree of likeness to cultural norms exhibited by individuals who are expected, as members of this society, to conform to those group patterns. In other words, the standardized test does not tell very much about the individual, *as an individual*, but rather how nearly he approximates to a normal performance of culturally prescribed tasks for which a more or less arbitrary, but internally consistent, scheme of quantitative ratings is utilized.<sup>6</sup> By the use of an all-over total figure for an individual, it becomes possible to assign numerical evaluations to individuals in various categories of achievement, skill, conformity, and so forth, such as accelerated, average, or retarded mentally; manual or verbal proficiency, etc. Having assigned him to a rank order in a group or class according to the standardized test, the individual is disposed of and adequately explained.<sup>7</sup> The history of the use of standardized tests shows how they are used to place individuals in various classifications that are convenient for administration, for remedial work and therapy, or for segregation for purposes of social control, with little or no concern about understanding the individual so classified or placed, or discovering his characteristics *as an individual*.

It would seem fair to say, therefore, that standardized tests offer procedures for rating individuals in terms of their socialization and how nearly they approximate to the acceptance and use of the culturally prescribed patterns of belief, action, and speech for which statistical norms can be calculated from actual observations of performance of *groups* of individuals, according to age, sex, etc.

In order to apply these and more recently developed quantitative methods to the study of personality it has been necessary to adopt a conception of the personality as an aggregation of discrete, measurable traits, factors, or other separable entities which are present in the individual in differing quantity and organized according to individual patterns. But since the personality is more than overt activity, some way of getting at the underlying personality is necessary. The need for quantitative data has led to the

5. Cf. Frank, L. K. Comments on the proposed standardization of the Rorschach method. *Rorschach Res. Exch.*, 1939, 3.

6. Cf. Kent, Grace H. Use and abuse of mental tests in clinical diagnosis. *Psychol. Rec.*, 1938, 2, 391-400.

7. Cf. Lewin, K. *A Dynamic Theory of Personality*. New York: McGraw-Hill, 1935. (Especially Chapter I on Aristotelian and Galilean modes of thought, and the class theory of investigation.)



use of the culturally standardized, socially sanctioned norms of speech and belief and attitudes in and through which the individual has been asked to express his personality, as in questionnaires, inventories, rating scales, etc.

If time allowed, it would be desirable to examine more fully the implications of this procedure which attempts to reveal the individuality of the person by using the social stereotypes of language and motives that necessarily subordinate individuality to social conformity, emphasizing likeness and uniformity of group patterns. This point becomes more significant when we recall that almost identical actions and speech may be used in extraordinarily different senses by each individual using them; while conversely, the widest diversity of action and speech may have almost identical sense and significance for different individuals exhibiting them. Moreover the conventional traits and motives and objectives derived from traditional concepts of human nature and conduct, carry meanings often alien to the investigator using them as data. Words are generalized symbols, are usually obscuring of, when not actually misleading about, the individual idiomatic personality using them.<sup>8</sup>

It should be further noted that many procedures for study of personality rely upon the subject's self-diagnosis and revelation of his *private world* of personal meanings and feelings which the social situation compels the individual to conceal, even if, as is unusual, he had any clear understanding of himself. When we ask an individual to tell what he believes or feels or to indicate in which categories he belongs, this social pressure to conform to the group norms operates to bias what he will say and presses him to fit himself into the categories of the inventory or questionnaire offered for self-diagnosis.<sup>9</sup> Moreover, as Henry A. Murray has observed, the most important things about an individual are what he cannot or will not say. The law has long recognized testimony as unreliable, to be accepted only after many checks and tests as formulated in the law of evidence.

At this point there may be a feeling of dismay, if not resentment, because the discussion has led to a seeming impasse, with no road open to study the personality by the accepted methods and procedures of present-day quantitative psychology. Moreover, the insistence upon the unique, idiomatic character of the personality appears to remove it from the area of scientific study conceived as a search for generalizations, uniformities, invariant relationships, etc. It is proposed, therefore, to discuss a few recent developments in scientific concepts and methods and the new problems they have raised in order to indicate a way out of this apparent impasse.

8. Cf. Willoughby, R. P., & Morse, Mary E. Spontaneous reactions to a personality inventory. *Amer. J. Orthopsychiat.*, 1936, 6, 562-575.

9. Cf. Vigotsky, L. S. Thought and speech. *Psychiatry*, 1939, 2, 29-54.

## B

It is appropriate to recall that the uniformity and laws of nature are statistical findings of the probable events and relationships that occur among an aggregate of events, the individuals of which are highly disorderly and unpredictable. Theoretical physics has adjusted itself to the conception of a universe that has statistical regularity and order, and individual disorder, in which the laws of aggregates are not observable in the activity of the individual making up these aggregates. Thus quantum physics and statistical mechanics and many other similar contrasts are accepted without anxiety about scientific respectability. The discrete individual event can be and is regarded as an individual to whom direct methods and measurements have only a limited applicability. We can therefore acknowledge an interest in the individual as a scientific problem and find some sanction for such an interest.

Another recent development is the concept of the *field* in physics and its use in biology. The field concept is significant here because it offers a way of conceiving this situation of an individual part and of the whole, which our older concepts have so confused and obscured.<sup>10</sup> Instead of a whole that dominates the parts, which have to be organized by some mysterious process into a whole, we begin to think of an aggregate of individuals which constitute, by their interaction, a field that operates to pattern these individuals. Parts are not separate, discrete, independent entities that get organized by the whole, nor is the whole a superior kind of entity with feudal power over its parts, e.g., a number of iron filings brought close to a magnet will arrange themselves in a pattern wherein each bit of iron is related to the other bits and the magnet and these relations constitute the whole; remove some bits and the pattern shifts as it does if we add more filings, or bits of another metal. Likewise, in a gas, the gas may be viewed as a field in which individual molecules, atoms, and electrons are patterned by the total interactions of all those parts into the group activity we call a gas. Ecology studies this interaction of various organizations in the circumscribed life zone or field which they constitute.<sup>11</sup>

This field concept is highly important because it leads to the general notion that any "entity" we single out for observation is participating in a field; any observation we make must be ordered to the field in which it is

10. Cf. Burr, H. S., & Northrop, F. S. C. An electro-dynamic theory of life. *Quart. Rev. Biol.*, 1935, 10, 322-333.

11. Cf. du Nouy, P. L. *Biological Time*. New York: Macmillan, 1937. (Other part-whole fields are a candle flame, a fountain jet, a stream of water, etc.)



made or as we say, every observation or measurement is relative to the frame of reference or field in which it occurs.

There are many other far-reaching shifts in concepts and methods that should be discussed here, but the foregoing will suffice to indicate that the study of an individual personality may be conceived as an approach to a somewhat disorderly and erratic activity, occurring in the field we call culture (i.e., the aggregate interaction of individuals whose behavior is patterned by participation in the aggregate). Moreover, the observations we make on the individual personality must be ordered to the field of that individual and his life space. We must also regard the individual himself as an aggregate of activities which pattern his parts and functions.

Here we must pause to point out that the older practice of creating entities out of data has created many problems that are unreal and irrelevant and so are insoluble. In by-gone years it was customary to treat data of temperature, light, magnetic activity, radiation, chemical activity, and so on as separate entities, independent of each other. But the more recent view is to see in these data evidences of energy transformations which are transmitted in different magnitudes, sequences, etc., and so appear as heat, light, magnetism, etc. This has relevance to the study of personality since it warns us against the practice of observing an individual's actions and then reifying these data into entities called traits (or some other discrete term), which we must then find some way of organizing into the living total personality who appears in experience as a unified organism.

With this background of larger, more general shifts in scientific procedures, let us examine some more specific developments that are relevant to our topic.

Within recent years new procedures have been developed for discovering not only the elements or parts composing the whole, but also the way those parts are arranged and organized in the whole, without disintegrating or destroying the whole. The X-rays are used, not merely for photographs or to show on a fluorescent screen what is otherwise invisible within an organism or any object, but also for diffraction analysis, in which the X-rays are patterned by the internal organization of any substance to show its molecular and atomic structure. Spectrographic analysis reveals the chemical components qualitatively, and now quantitatively, and in what compounds, by the way light is distributed along a continuous band of coarse and fine spectral lines, each of which reveals a different element or isotope. The mass spectroscopy offers another exceedingly delicate method for determining the composition of any substance that gives off radiations whereby the electrons or their rate of travel can be measured and the composition of the substance inferred.

X-rays, however, are only one of the newer methods whereby any complex may be made to reveal its components and its organization, often quantitatively, when approached by an appropriate procedure. Recently, it has been found that the chemical composition of various substances, especially proteins, can be ascertained by the reflection of a beam of light upon a thin monomolecular film of the protein substance spread on a film of oil on water over a metallic surface. Again, it has been found that metallic ores and coal may be analyzed, i.e., be made to reveal their chemical composition and other properties by the "angle of wetability," the angle of reflection, or the color of the light reflected from a liquid film that adheres to the surface of the unknown material.

Polarized light has also become an instrument for revealing the chemical composition of substances without resort to the usual methods of disintegration or chemical analysis. Electrical currents may also be passed through substances, gaseous, liquid, or solid, and used to discover what they contain and in what form. Indeed, it is not unwarranted to say that these indirect methods that permit discovery of the composition and organization of substances, complexes, and organisms, seem likely to become the method of choice over the older destructive analytical procedures, because these methods do not destroy or disturb the substance or living organism being studied.

In this connection reference should also be made to the development of biological assays, whereby a living organism, plant or animal, is used for assaying the composition of various substances and compounds and determining their potency, such as vitamins, hormones, viruses, drugs, radiation, light, magnetism, and electrical currents (including electrophoresis for separating, without injury or change, the different subvarieties of any group of cells, chemical substances, etc.). In these procedures the response of the living organism is utilized as an indicator, if not an actual measurement, of that about which data are sought, as well as the state, condition, maturation, etc., of the organism being tested. It is appropriate to note also that physicists are using such devices as the Wilson cloud chamber and the Geiger counter to obtain data on the *individual* electrical particle, which reveals its presence and energy by the path traced in water vapor, or by activation of the counter, although never itself observable or directly measurable.

These methodological procedures are being refined and extended because they offer possibilities for ascertaining what is either unknowable by other means or is undeterminable because the older analytic methods destroyed part or all of that which was to be studied. They are being accepted as valid and credible, primarily because they are more congruous with the search for undivided totalities and functioning organisms and are more



productive of the data on organization on which present-day research problems are focused. They are also expressive of the recent concepts of whole-and-parts and their interrelations, which no longer invoke the notion of parts as discrete entities upon which an organization is imposed by a superior whole, but rather employ the concept of the field. Finally, they offer possibilities for studying the specific, differentiated individuality of organized structures and particulate events which are ignored or obscured by the older quantitative determinations of aggregates.

Since the threshold task in any scientific endeavor is to establish the meanings and significances of the data obtained by any method of observation and measurement, it should be noted that these indirect methods for revealing the composition and organization of substances and structures rely upon experimental and genetic procedures to establish reliability and validity, not statistical procedures. That is to say, these newer procedures establish the meaning of any datum by employing the procedure upon a substance or structure of known composition, often made to order, so that it is possible to affirm that the resulting bending, patterning, arrangement of light, radiation, and so on, are reliable and valid indicators of the substance or structure when found in an unknown composition. These methods for establishing reliability and validity are therefore genetic in the sense of observing or tracing the origin and development of what is to be tested so that its presence or operation may be historically established: they are also dependent upon the concurrent use of other procedures which will yield consistent data on the same composition which therefore are validated by such internal consistency and congruity of findings.

Psychology developed the statistical procedures for establishing reliability and validity because the only data available were the single observations or measurements taken at one time on each subject. Since no other data were available on the prior history and development of the subjects, reliability had to be determined by statistical manipulation of these test data themselves; also, since no other data were available on the subject's other functions and activities only statistical validity could be established. It would appear that these tests of reliability and validity devised to meet the difficulty presented by absence of other data now act as barriers to the use of any other procedures for personality study in which reliability and validity for each subject is tested through these other nonstatistical methods.

Methods of *temporal validation* offer great promise because they permit testing of the validity of data for a *specific subject* over a period of time, and the method of congruity among data obtained by different procedures from the same subject offer large possibilities for testing the reliability of any

data for a *specific subject*.<sup>12</sup> It is appropriate to recall here that the accepted methods for testing reliability and validity of tests, inventories, etc., offer indices only for the *group*, not for any individual subject in that *group*.

We may therefore view the problem of personality in terms of these more recent ideas and conceptions and consider the application of these indirect procedures for revealing the composition and organization of substances and energy complexes.

As indicated earlier the personality may be viewed as a dynamic process of organizing experience, of "structuralizing the life space" (Lewin) according to the unique individual's *private world*. This conception may be made precise and operational by seeing the individual and his changing environment as a series of fields which arise through the interaction of the individual personality with his selective awareness, patterned responses, and idiomatic feelings, with the environmental situations of objects, events, and other persons. A field organization or configuration arises out of this interaction wherein, as suggested, the personality distorts the situation, so far as it is amenable, into the configurations of its *private world*, but has to adjust to the situation in so far as it resists such distortion and imposes its necessities upon the personality. What we have called personality and fumblingly have tried to formulate as the total responses of the whole individual and similar additive conceptions becomes more understandable and approachable for investigation when conceived as the living process in this field created by the individual and the environing situation.

The objective world of objects, organisms, and events likewise may be seen as fields of interacting object-situations, upon which cultural patterns operate in the conduct of human beings who, by very reason of behaving in these learned patterns, create the cultural fields of interacting human conduct. What is highly important to note is that every observation made must be ordered—given its quantitative and qualitative interpretation—to the field in which it occurs, so that the idea of pure objectivity becomes meaningless and sterile if it implies data not biased, influenced, relative to the field in which observed. Likewise the conception of a stimulus that may be described and measured apart from the field and the organism in that field is untenable.<sup>13</sup> The "same" stimulus will differ in every field, and for every field and for every organism which selectively creates its own stimuli in

12. Cf. Bateson, G. *Naven*. Cambridge: Cambridge Univ. Press, 1936, in which appears a discussion of diachronic and synchronic procedures.

13. Cf. Vigotsky, L. S. *Thought and speech*. *Psychiatry*, 1939, 2, 29–54. "The investigator who uses such methods may be compared to a man, who, in order to explain why water extinguishes fire, analyzes the water into oxygen and hydrogen and is surprised to find that oxygen helps the process of burning and hydrogen itself burns. This method of analyzing a whole into elements is not a true analysis which can be used to solve concrete problems" (p. 29).



each situation. Indeed, this dynamic conception of the personality as a process implies that there are no stimuli to conduct (as distinct from physical and physiological impacts) except in so far as the individual personality selectively constitutes them and responds to them in its idiosyncratic patterns. In other words the stimuli are functions of the field created by the individual interacting with the situation.

Thus the movement in various areas of scientific work is toward recognition of the field concept and the devising of methods that will record not merely data but the fields in which those data have been observed and find their significance. Those who are appalled by the seeming anarchy thus threatening scientific work may be reminded that the present-day standards of scientific work and of methods are part of a development that will inevitably make today's ideas and procedures obsolete. It is well to recall how proud (justly so) chemistry was to achieve quantitative determinations of the composition of substances and now, how crude those early quantitative methods and findings now appear, when they now are seeking to find out not merely what and how much, but the spatial arrangement of the constituents as in stereochemistry where the same atoms in the same quantity produce different substances according to their spatial arrangement. It is likewise worth recalling, that about 1900, young physicists could find no problems except the more precise measurement of the pressure, temperature, etc., of a gas and were content with such crude quantitative findings. Furthermore, biologists today are accepting as commonplace that the same nutritive components, amino-acids, carbohydrates, fats, minerals, and vitamins are selectively digested, assimilated, and metabolized in different ways by each species and by each individual. Moreover, it is conceded that the proteins of each species are different as are those of each individual with the possibility of an almost unlimited number of different protein molecules, in which the same basic elements are organized into unique spatial-temporal configurations appropriate to the organic field of the individual organism.<sup>14</sup>

## C

Coming directly to the topic of projective methods for personality study,<sup>15</sup> we may say that the dynamic conception of personality as a process

14. The concepts of individuality and of individuation are being used by biologists because they find themselves confronted with individual organic activities and idiomatic processes. Cf. Blumenthal, H. T. Effects of organismal differentials on the distribution of leukocytes in the circulating blood. *Arch. Path.*, 1939, 27, 510-545. Cf. Coghill, G. E. Individuation versus integration in the development of behavior. *J. Gen. Psychol.*, 1930, 3, 431-435. Cf. Coghill, G. E. Integration and motivation of behavior as problems of growth. *J. Genet. Psychol.*, 1936, 48, 3-19.

15. References to the projective techniques discussed in this section appear in the references at the end of the chapter.

of organizing experience and structuralizing life space in a field, leads to the problem of how we can reveal the way an individual personality organizes experience, in order to disclose or at least gain insight into that individual's *private world* of meanings, significances, patterns, and feelings.

Such a problem is similar to those discussed earlier where indirect methods are used to elicit the pattern of internal organization and of composition without disintegrating or distorting the subject, which is made to bend, deflect, distort, organize, or otherwise pattern part or all of the field in which it is placed—e.g., light and X-rays. In similar fashion we may approach the personality and induce the individual to reveal his way of organizing experience by giving him a field (objects, materials, experiences) with relatively little structure and cultural patterning so that the personality can project upon that plastic field his way of seeing life, his meanings, significances, patterns, and especially his feelings. Thus we elicit a projection of the individual personality's *private world* because he has to organize the field, interpret the material and react affectively to it. More specifically, a projection method for study of personality involves the presentation of a stimulus-situation designed or chosen because it will mean to the subject, not what the experimenter has arbitrarily decided it should mean (as in most psychological experiments using standardized stimuli in order to be "objective"), but rather whatever it must mean to the personality who gives it, or imposes upon it, his private, idiosyncratic meaning and organization. The subject then will respond to *his* meaning of the presented stimulus-situation by some form of action and feeling that is expressive of his personality. Such responses may be *constitutive* as when the subject imposes a structure or form or configuration (Gestalt) upon an amorphous, plastic, unstructured substance such as clay, finger paints, or upon partially structured and semi-organized fields like the Rorschach cards; or they may be *interpretive* as when the subject tells what a stimulus-situation, like a picture, means to him; or they may be *cathartic* as when the subject discharges affect or feeling upon the stimulus-situation and finds an emotional release that is revealing of his affective reactions toward life situations represented by the stimulus-situation, as when he plays with clay or toys. Other expressions may be *constructive* organizations wherein the subject builds in accordance with the materials offered but reveals in the pattern of his building some of the organizing conceptions of his life at that period, as in block-building.

The important and determining process is the subject's personality which operates upon the stimulus-situation as if it had a wholly private significance for him alone or an entirely plastic character which made it yield to the subject's control. This indicates that, as suggested earlier, a personality is the way an individual organizes and patterns life situations and



effectively responds to them, "structuralizes his life space," so that by projective methods we are evoking the very process of personality as it has developed to that moment.<sup>16</sup> Since the way an individual organizes and patterns life situations, imposes his *private world* of meanings and affectively reacts upon the environing world of situations and other persons and strives to maintain his personal version against the coercion or obstruction of others, it is evident that personality is a persistent way of living and feeling that, despite change of tools, implements, and organic growth and maturation will appear continuously and true to pattern.

When we scrutinize the actual procedures that may be called projective methods we find a wide variety of techniques and materials being employed for the same general purpose, to obtain from the subject, "what he cannot or will not say," frequently because he does not know himself and is not aware what he is revealing about himself through his projections.

In the following statement no attempt has been made to provide a complete review of all the projective methods now being used, since such a canvass would be beyond the present writer's competence and intention. Only a few illustrations of projective methods are offered to show their variety and their scope, in the hope of enlisting further interest in and creating a better understanding of, their characteristics and advantages.<sup>17</sup>

The Rorschach ink blots, to which the subject responds by saying what he "sees" in each of a number of different blots, are perhaps the most widely known of these procedures. They have been utilized in Europe and in the United States, frequently in connection with psychiatric clinics and hospitals, for revealing the personality configurations and have been found of increasing value. Insofar as life histories and psychiatric and psychoanalytic studies of the subjects who have had the Rorschach diagnosis are available, the ink blot interpretations are being increasingly validated by these clinical findings. Such comparative findings are of the greatest importance because they mutually reinforce each other and reveal the consistency or any conflicts in the different interpretations and diagnosis of a personality.

Another similar procedure is the *Cloud Picture* method, developed by Wilhelm Stern, to evoke projections from a subject upon more amorphous grounds, with advantages, he believed, over the Rorschach blots. The more amorphous or unstructured the ground, the greater the sensitivity of the procedure which however loses in precision as in most instruments. Hence

16. Cf. Dunbar, H. F. *Emotions and Bodily Changes*. New York: Columbia Univ. Press, 1938, (2nd ed.) An individual may express his feelings, otherwise blocked, in illness or physiological dysfunctions.

17. Cf. Horowitz, Ruth, & Murphy, Lois B. Projective methods in the psychological study of children. *J. Exper. Educ.*, 1938, 7, 133-140, for further discussion of different procedures and their use.

the Rorschach may be less sensitive than *Cloud Pictures* or clay but more precise and definite. Both the ink blots and the *Cloud Pictures* offer a ground upon which the subject must impose or project whatever configurational patterns he "sees" therein, because he can see only what he personally looks for or "perceives" in that ground. The separate details of the responses, however, are significant only in the context of the total response to each blot and are meaningful only for each subject. This does not imply an absence of recurrent forms and meanings from one subject to another but rather that the same letters of the conventionalized alphabet may recur in many different words and the same words may be utilized in a great variety of sentences to convey an extraordinary diversity of statements, which must be understood within the context in which they occur and with reference to the particular speaker who is using them on that occasion.<sup>18</sup>

Play techniques are being increasingly employed for clinical diagnosis and for investigation of the personality development of children. As materials almost any kind of toy or plaything or plain wooden building blocks may be presented to the subject for free play or for performance of some designated action, such as building a house, sorting into groups, setting the stage for a play or otherwise organizing the play materials into some configuration which expresses for the subject an affectively significant pattern. In children, it must be remembered there are fewer disguises and defenses available to hide behind and there is less sophisticated awareness of how much is being revealed in play. The investigator does not set a task and rate the performance in terms of skill or other scale of achievement, since the intention is to elicit the subject's way of "organizing his life space" in whatever manner he finds appropriate. Hence every performance is significant, apart from the excellence of the play construction or activity, and is to be interpreted, rather than rated, for its revelation of how the subject sees and feels his life situations that are portrayed in the play constructions and sequences. The question of how to decide whether a particular activity is or is not meaningful is to be decided, not by its frequency or so-called objective criteria, but by the total play configuration of that particular subject who, it is assumed, performs that particular action or uses that specific construction, as an expression of his way of seeing and feeling and reacting to life, i.e., of his personality. But the degree of relevance is to be found in the context, in what precedes and what follows and in the intensity of feelings expressed. If these criteria appear tenuous and subjective and lacking in credibility, then objections may be made to the use of various methods for

18. Cf. Since each personality must use socially prescribed cultural patterns for his conduct and communications he will exhibit many recurrent uniformities but these are significant only for revealing the patterns or organizations or configurations which the personality uses to structuralize his life space.



discovering the composition and structure of an unknown substance through which light, electric current, or radiations are passed, to give patterned arrangements or a spectrum photograph in which the position, number, intensity of lines and the coarse and fine structure indicate what the unknown substance is composed of, how organized internally, and so on. Personality studies by projective methods have not, of course, been as extensively studied nor have the patterns used by subjects been so well explored. The important point is that the way is open to the development of something similar to spectroscopic and diffraction methods for investigation of personality.

If the foregoing appears far-fetched it may be recalled that the lines on the spectroscopic plate were established, not by statistical procedures, but by experimental procedures through which a known chemically tested substance was spectroscopically tested so that its identifying line could be precisely located and thereafter confidently named. In much the same fashion it is being established that a child who is known to be undergoing an affective experience will express that feeling in a play configuration that can be so recognized. Thus, children who have lost a beloved parent or nurse, who have been made anxious by toilet training, are insecure and hostile because of sibling rivalry, etc., will exhibit those feelings in their play configurations. Experimentally produced personality disturbances can be established and their severity investigated by subsequent play forms and expressions. Moreover, the insights derived from play configurations yield interpretations that are not only therapeutically effective but often predictive of what a child will show in the near future.

Not only play toys and objects are utilized but also various plastic amorphous materials such as modeling clay, flour and water, mud and similar substances of a consistency that permits the subject to handle freely and manipulate into various objects. In these play situations the subject often finds a catharsis, expressing affects that might otherwise be repressed or disguised, or symbolically releasing resentments and hostility that have been long overlaid by conventionally good conduct. Dolls, capable of being dismembered, can be used to evoke repressed hostility and aggression against parents and siblings. Dramatic stage play with toy figures and settings have also provided occasions in which a subject not only revealed his personality difficulties but also worked out many of his emotional problems. Clay figures are modeled by child patients in which they express many of their acute anxieties and distortions. Reference should be made to eidetic imagery, which, as Walther Jaensch in his constitutional studies has shown, indicates one aspect of the subject's way of expressing what enters into his personality make-up or way of organizing his life space.

Artistic media offer another series of rich opportunities for projective methods in studying personality. Finger-painting has given many insights into child personality make-up and perplexities. Painting has been found very fruitful for study of personality make-up and emotional disturbances. Other clinical uses of painting have been reported that indicate the way paintings and drawings supplement the clinician's interviews and evoke responses that are exceedingly revealing, often more so than the verbal responses. Puppet shows elicit responses from child patients that are both diagnostic and therapeutic because the intensity of the dramatic experience arouses the child to a vehement expression of his feelings toward authority and toward parents and of his repressed desires to hurt others. Roles have been assigned to individuals who are then asked to act out those roles impromptu, thereby revealing how tangled and repressed his or her feelings are and how release of pent-up emotion leads to insight into one's personality difficulties. Dramatic teachers are finding clues to personality in the way individuals portray the characters assigned them in a play. Music offers similar and often more potent possibilities for expression of affects that are revealing of the personality. It is interesting to note that as psychotherapy proceeds to free the patient, his art expressions, painting, modeling, music, and dramatic rendition become freer and more integrated.

As the foregoing indicates, the individual rarely has any understanding of himself or awareness of what his activities signify. In the Thematic Perception methods this unawareness offers an opportunity to elicit highly significant projections from subjects who are asked to write or tell stories about a series of pictures showing individuals with whom they can identify themselves and others of immediate personal concern. Likewise the subjects project many aspects of their personality in the completion of stories and of sentences, in making up analogies, sorting out and grouping objects, such as toys, and similar procedures in which the subject reveals "what he cannot or will not say."

Expressive movements, especially handwriting, offer another approach to the understanding of the personality who reveals so much of his characteristic way of viewing life in his habitual gestures and motor patterns, facial expressions, posture and gait. These leads to the study of personality have been rejected by many psychologists because they do not meet psychometric requirements for validity and reliability, but they are being employed in association with clinical and other studies of personality where they are finding increasing validation in the consistency of results for the same subject when independently assayed by each of these procedures. In this group of methods should be included observations on tics of all kinds and dancing as indications of tension, anxiety or other partially repressed feelings.



If we will face the problem of personality, in its full complexity, as an active dynamic process to be studied as a *process* rather than as entity or aggregate of traits, factors, or as a static organization, then these projective methods offer many advantages for obtaining data on the process of organizing experience which is peculiar to each personality and has a life career. Moreover, the projective methods offer possibilities for utilizing the available insights into personality which the prevailing quantitative procedures seem deliberately to reject.

Here again it may be re-emphasized that the study of personality is not a task of measuring separate variables on a large group of individuals at one moment in their lives and then seeking, by statistical methods, to measure correlations, nor is it a problem of teasing out and establishing the quantitative value of several factors.<sup>19</sup> Rather the task calls for the application of a multiplicity of methods and procedures which will reveal the many facets of the personality and show how the individual "structuralizes his life space" or organizes experience to meet his personal needs in various media. If it appears that the subject projects similar *patterns* or *configurations* upon widely different materials and reveals in his life history the sequence of experiences that make those projections psychologically meaningful for his personality, then the procedures may be judged sufficiently valid to warrant further experimentation and refinement. In undertaking such explorations the experimenter and clinicians may find reassurance and support in the realization that they are utilizing concepts and methods that are receiving increasing recognition and approval in scientific work that is today proving most fruitful.

19. Cf. Jersild, A. T., & Fite, Mary D. The influence of nursery school experience on children's social adjustments. *Child Devel. Monog.*, No. 25, 1939. See especially page 102.

## REFERENCES

1. ABEL, THEODORA M. Free designs of limited scope as a personality index. *Charac. & Person.*, 1938, 7, 50-62.
2. ACKERMAN, N. W., with the technical assistance of VIRGINIA REHKOPF. Constructive and destructive tendencies in children. *Amer. J. Orthopsychiat.*, 1937, 7, 301-319.
3. ALLPORT, G. W., & VERNON, P. E. *Studies in Expressive Movement*. New York: Macmillan, 1933. P. 269.
4. ANDERSON, H. H. Domination and integration in the social behavior of young children in an experimental play situation. *Genet. Psychol. Monog.*, 1937, 19, 343-408.

5. BARKER, R. G. The effect of frustration upon cognitive ability. *Charac. & Person.*, 1938, 7, 145-150.
6. BARKER, R. G., DEMBO, T., & LEWIN, K. Experiments in frustration and regression studies in topological and vector psychology. *Iowa Child Wel. Res. St. Monog.*, 1939.
7. BECK, S. J. Autism in Rorschach scoring: A feeling comment. *Charac. & Person.* (News and Notes), 1936, 5, 83-85.
8. ———. Psychological processes in Rorschach findings. *J. Abn. & Soc. Psychol.*, 1937, 31, 482-488.
9. ———. Introduction to the Rorschach method. *Amer. Orthopsychiat. Assoc. Monog.*, No. 1, 1937.
10. ———. Personality structure in schizophrenia. *Nerv. & Ment. Dis. Monog.*, 1938, No. 63, p. 88.
11. BENDER, LAURETTA, & WOLTMANN, A. The use of puppet shows as a psychotherapeutic method for behavior problems in children. *Amer. J. Orthopsychiat.*, 1936, 6, 341-354.
12. BENDER, LAURETTA, KEISER, S., & SCHILDER, P. Studies in aggressiveness. *Genet. Psychol. Monog.*, 1936, 18, 357-564.
13. BENDER, LAURETTA, & SCHILDER, P. Form as a principle in the play of children. *J. Genet. Psychol.*, 1936, 49, 254-261.
14. BENDER, LAURETTA, & WOLTMANN, A. Puppetry as a psychotherapeutic measure with problem children. *Monthly Bull., N.Y. State Assoc. Occup. Therap.*, 1937, 7, 1-7.
15. BENDER, LAURETTA. Art and therapy in the mental disturbances of children. *J. Nerv. & Ment. Dis.*, 1937, 86, 229-238.
16. ———. Group activities on a children's ward as methods of psychotherapy. *Amer. J. Psychiat.*, 1937, 93, 1151-1173.
17. BENDER, LAURETTA, & WOLTMANN, A. The use of plastic material as a psychiatric approach to emotional problems in children. *Amer. J. Orthopsychiat.*, 1937, 7, 283-300.
18. BENDER, LAURETTA. A visual motor gestalt test and its clinical use. *Amer. Orthopsychiat. Assoc. Monog.*, 1938, No. 3, p. 176.
19. BOOTH, G. C. Personality and chronic arthritis. *J. Nerv. & Ment. Dis.*, 1937, 85, 637-662.
20. ———. The use of graphology in medicine. *J. Nerv. & Ment. Dis.*, 1937, 86, 674-679.
21. ———. Objective techniques in personality testing. *Arch. Neur. & Psychiat.*, 1939.
22. CAMERON, N. Individual and social factors in the development of graphic symbolization. *J. of Psychol.*, 1938, 5, 165-183.
23. ———. Functional immaturity in the symbolization of scientifically trained adults. *J. of Psychol.*, 1938, 6, 161-175.
24. ———. Reasoning, regression, and communication in schizophrenics. *Psychol. Monog.*, 1938, 50, 1-34.



25. ———. Deterioration and regression in schizophrenic thinking. *J. Abn. & Soc. Psychol.*, 1939, 34, 265-270.
26. CONN, J. H. A psychiatric study of car sickness in children. *Amer. J. Orthopsychiat.*, 1938, 8, 130-141.
27. CURRAN, F. F. The drama as a therapeutic measure in adolescents. *Amer. J. Orthopsychiat.*, 1939, 9, 215-231.
28. DESPERT, J. L., & POTTER, H. W. The story, a form of directed phantasy. *Psychiat. Quart.*, 1936, 10, 619-638.
29. ERIKSON, E. H. Configurations in play: Clinical notes. *Psychoanal. Quart.*, 1937, 6, 139-214.
30. FITE, MARY D. Aggressive behavior in young children and children's attitudes toward aggression. *Genet. Psychol. Monog.*, 1939.
31. GERARD, MARGARET W. Case for discussion at the 1938 symposium. *Amer. J. Orthopsychiat.*, 1938, 8, 1-8.
32. GITELSON, M. (Chairman), *et al.* Section on "play therapy," 1938. *Amer. J. Orthopsychiat.*, 1938, 8, 499-524.
33. GITELSON, M. Clinical experience with play therapy. *Amer. J. Orthopsychiat.*, 1938, 8, 466-478.
34. GRIFFITHS, R. *Imagination in Young Children*. London: Kegan Paul, 1936.
35. HANFMANN, EUGENIA. Social structure of a group of kindergarten children. *Amer. J. Orthopsychiat.*, 1935, 5, 407-410.
36. HANFMANN, EUGENIA, & KASANIN, J. A method for the study of concept formation. *J. of Psychol.*, 1937 3, 521-540.
37. ———. Disturbances in concept formation in schizophrenia. *Arch. Neur. & Psychiat.*, 1938, 40, 1276-1282.
38. HANFMANN, EUGENIA. Analysis of the thinking disorder in a case of schizophrenia. *Arch. Neur. & Psychiat.*, 1939, 41, 568-579.
39. HERTZ, MARGUERITE R. The method of administration of the Rorschach ink blot test. *Child Devel.*, 1936, 7, 237-254.
40. HERTZ, MARGUERITE R., & RUBENSTEIN, B. B. A comparison of three "blind" Rorschach analyses. *Amer. J. Orthopsychiat.*, 1939, 9, 295-314.
41. HOLMER, P. The use of the play situation as an aid to diagnosis. *American J. Orthopsychiat.*, 1937, 7, 523-531.
42. HOROWITZ, RUTH E., & MURPHY, LOIS B. Projective methods in the psychological study of children. *J. Exper. Educ.*, 1938, 7, 133-140.
43. HOROWITZ, RUTH E. Racial aspects of self-identification in nursery school children. *J. of Psychol.*, 1939, 7, 91-99.
44. HUNTER, MARY. The practical value of the Rorschach test in a psychological clinic. *Amer. J. Orthopsychiat.*, 1939, 9, 287-294.
45. JAENSCH, E. R. *Eidetic Imagery and Typological Methods of Investigation*. New York: Harcourt, Brace, 1930.
46. KASANIN, J., & HANFMANN, EUGENIA. An experimental study of concept formation in schizophrenia: I. Quantitative analysis of the results. *Amer. J. Psychiat.*, 1938, 95, 35-48.

47. KELLEY, D. M., & KLOFFER, B. Application of the Rorschach method to research in schizophrenia. *Rorschach Res. Exch.*, 1939, 3, 55-66.
48. KLOFFER, B. *Rorschach Research Exchange*. September, 1936, to date.
49. KLÜVER, H. The Eidetic Child, in *Handbook of Child Psychology*. Worcester: Clark Univ. Press, 1931.
50. LEVY, D. M. Use of play technique as experimental procedure. *Amer. J. Orthopsychiat.*, 1933, 3, 266-277.
51. ———. Hostility patterns in sibling rivalry experiments. *Amer. J. Orthopsychiat.*, 1936, 6, 183-257.
52. ———. "Release therapy" in young children. *Psychiatry*, 1938, 1, 387-390.
53. LEVY, J. The use of art techniques in treatment of children's behavior problems. *Proc. Amer. Assoc. Ment. Def.*, 1934, 58, 258-260.
54. ———. The active use of phantasy in treatment of children's behavior problems. (Unpublished paper presented at a meeting of the American Psychiatric Association.)
55. LEWIN, K. Environmental forces. *Handbook of Child Psychology*. Worcester: Clark Univ. Press, 1933.
56. ———. *A Dynamic Theory of Personality*. New York: McGraw-Hill, 1935. Pp. 286.
57. ———. *Principles of Topological Psychology*. New York: McGraw-Hill, 1936. Pp. 231.
58. ———. Psychoanalysis and topological psychology. *Bull. Menninger Clin.*, 1937, 1, 202-211.
59. LISS, E. Play techniques in child analysis. *Amer. J. Orthopsychiat.*, 1936, 6, 17-22.
60. ———. The graphic arts. *Amer. J. Orthopsychiat.*, 1938, 8, 95-99.
61. LOWENFELD, V. *The Nature of Creative Activity*. London: Kegan Paul, 1938.
62. MASSERMAN, J. H., & BALKEN, EVA R. The clinical application of phantasy studies. *J. of Psychol.*, 1938, 6, 81-88.
63. MORENO, J. L. Who shall survive? *Nerv. & Ment. Dis. Monog.*, 1934, No. 58.
64. MORENO, J. L., & JENNINGS, H. Spontaneity training, a method of personality development. *Sociomet. Rev.*, 1936.
65. MORENO, J. L. Psychodramatic shock therapy—A sociometric approach to the problem of mental disorders. *Sociometry*, 1939, 2, 1-30.
66. ———. Creativity and cultural conserves—with special reference to musical expression. *Sociometry*, 1939, 2, 1-36.
67. MORGAN, CHRISTINE D., & MURRAY, H. A. Method for investigating fantasies—the thematic apperception test. *Arch. Neur. & Psychiat.*, 1935, 34, 289-306.
68. MURPHY, LOIS B. *Social Behavior and Child Personality. An Exploratory Study of Some Roots of Sympathy*. New York: Columbia Univ. Press, 1937. P. 325.
69. MURRAY, H. A., et al. Papers in *J. Soc. Psychol.*, 1933, 4; *J. Abn. & Soc. Psychol.*, 1934, 28; *J. of Psychol.*, 1937, 3, 27-42.
70. MURRAY, H. A. Techniques for a systematic investigation of fantasy. *J. of Psychol.*, 1937, 3, 115-143.
71. MURRAY, H. A., et al. *Explorations in Personality*. New York: Oxford Univ. Press, 1938.



72. NEWMAN, S., & MATHER, VERA G. Analysis of spoken language of patients with affective disorders. *Amer. J. Psychiat.*, 1938, 94, 913-942.
73. NEWMAN, S. Personal symbolism in language patterns. *Psychiatry*, 1939, 2, 177-184.
74. PIOTROWSKI, Z. The methodological aspects of the Rorschach personality method. *Kwart. Psychol.*, at Poznan, 1937, 9, 29.
75. ———. The M, FM, and m responses as indicators of changes in personality. *Rorschach Res. Exch.*, 1937, 1, 148-156.
76. PLANT, J. S. Personality and the culture pattern. *J. Soc. Philos.*, 1938, 3, 126-142.
77. PORTER, E. L. H. Factors in the fluctuation of fifteen ambiguous phenomena. *Psychol. Rec.*, 1937, 2, 231-253.
78. ROSENZWEIG, S. & SHAKOW, D. Play technique in schizophrenia and other psychoses: I. Rationale; II. An experimental study of schizophrenic constructions with play materials. *Amer. J. Orthopsychiat.*, 1937, 7, 32-35, 36-47.
79. SAPIR, E. The emergence of the concept of personality in a study of culture. *J. Soc. Psychol.*, 1934, 5, 408-415.
80. SENDER, SADIE, & KLOPPER, B. Application of the Rorschach test to child behavior problems as facilitated by a refinement of the scoring method. *Rorschach Res. Exch.*, 1936. Issue No. 1, 1-17.
81. SHAW, R. F. *Finger Painting*. Boston: Little Brown, 1934. P. 232.
82. STEIN-LEWINSON, THEA. An introduction to the graphology of Ludwig Klages. *Charac. & Person.*, 1933, 6, 163-177.
83. TROUP, EVELYN. A comparative study by means of the Rorschach method of personality development in twenty pairs of identical twins. *Genet. Psychol. Monog.*, 1938, 20, 465-556.
84. VAUGHN, J., & KRUG, OTHILDA. The analytic character of the Rorschach ink blot test. *Amer. J. Orthopsychiat.*, 1938, 8, 220-229.
85. WERNER, H. William Stern's personalistics and psychology of personality. *Charac. & Person.*, 1938, 7, 109-125.

Helen Sargent

## PROJECTIVE METHODS

*Their Origins, Theory, and Application*

*in Personality Research*

PROJECTIVE METHODS were in use prior to 1939, but were not designated as such until after that date which marks the introduction of the term in an article by L. K. Frank (5). Since then, the concept has been specifically employed in an increasing number of titles in the *Psychological Abstracts*, and has become common usage in the literature of personality research. A lively experimental attack utilizing the projective approach has grown up in child psychology, psychopathology, and personality.

Fairly representative of definitions usually offered is the following:

A projective method for the study of personality involves the presentation of a stimulus situation designed or chosen because it will mean to the subject not what the experimenter has arbitrarily decided it should mean (as in most psychological experiments using standardized stimuli in order to be "objective") but rather whatever it must mean to the personality who gives it, or imposes upon it, his private, idiosyncratic meaning and organization (5, p. 403).

Reprinted from *Psychol. Bull.*, 1945, 42, 257-293 by permission of the American Psychological Association and the estate of the author.



These methods which Frank described, and for which he furnished a rationale as well as a name, are in no sense a new discovery, although their current popularity is in part derived from a research atmosphere peculiarly suited to their rapid growth in the past five years. The very wording of the above definition implies a controversy: it presents projective techniques not only as an addition to our present stock of instruments; it also implies that they are set up in opposition to something. In order to understand either their promise, or the obstacles which stand in the way of their unqualified welcome in scientific circles, it will be necessary to examine their historical roots as well as the contemporary theoretical climate in which they flourish; to survey the problems to which they have been applied and the results obtained; and to study certain methodological difficulties which beset them. Furthermore, it will be necessary to hold a patient hearing of somewhat repetitive controversial views.

## *Background*

The developing family of projective methods might be regarded as the legitimate children of two parents: a brilliant and daring mother, psychiatry, and an equally intelligent but more conservative father, academic psychology. The five-year-old offspring partake of the characteristics of both forebears; they have a promising future, but have not yet overcome insecurity engendered by the fact that each parent is inclined to berate them for faults presumably inherited from the other.

Putting metaphor aside, we may discern in the context which surrounds projective techniques, three major theoretical trends which have contributed to a general point of view, and four lines of research more or less closely related to projective experimentation. The most important theoretical influences include psychoanalysis, *global* theory, and certain developments in twentieth-century general science. Relevant research includes studies in imagination and phantasy, the word-association method, investigations of language, and the development of methods for the use and interpretation of *personal documents*.

## THEORETICAL CLIMATE

1. *Psychoanalysis*. The term *projection* was first used by Freud to describe one of the unconscious minor *mechanisms* of conflict solution (230). The ego, unable to accept in itself certain thoughts, wishes, or characteristics, attributes these to environmental objects or to persons (222, p. 75). Sears, quoting a somewhat elaborate definition by Healy, Bronner and Bowers,

which expresses a similar connotation, prefers to withdraw the term from its *metapsychological setting* and to define it as follows:

A wish, attitude, or habit-hierarchy which is not compatible with other attitudes or habits of an individual may be attributed by that individual to other persons rather than to himself, providing he lacks insight into the fact that he himself possesses the trait in question. This process of attribution is unconscious, i.e., the subject does not give any verbal evidence that he knows his perception is false (19, p. 561).

Sears has also called attention to a distinction between the above use of the term projection as applied to the basic paranoid mechanism, and its usage with reference to projective techniques. He states that "in the latter case the implication is that the motivational and organizational properties of a personality influence the perceptual and judgmental processes" (262, p. 121). Important as it is to differentiate these two connotations, there appears to be reason rather than confusion in the dual application of the word. The alleged defensive reactions of the ego are subject to observation in the selective effect upon perception, and in diverse expressive responses; conversely the motivations which are assumed to be operative are subject to explanation in terms of ego defense. It may be said that the noun *projection* describes one type of defense, and that the adjective *projective* applies more broadly to the observable effects of this and other psychic processes, and to the methods used to elicit and study them.

That the mechanism of projection is one of the most readily understood and accepted of any in the Freudian scheme is demonstrated by its easy translation into the idiom of a child, or of a schizophrenic. Feigenbaum reports that a little girl to whom a paranoid acquaintance had been described as "hating people" replied promptly: "I know why he hates them! It's like when Mother wants to go to the toilet, she asks me if I have to" (4, p. 305). A schizophrenic, one of Balken's subjects, remarked: "If I refuse to recognize it, it is not for me" (144, p. 249). The fields of art and literature also provide an almost unexplored territory for the study of projection as a psychological phenomenon. In fiction and in poetry it is possible to trace not only the projections of the writer, but to discover subtle techniques used by authors (and by musicians and artists as well) to provide a medium in which others may project and enjoy release. Dean Addison Hibbard of Northwestern University has for a number of years assigned items from the personal columns in English newspapers to students as a starting point for compositions, and reports that he has regularly noted the inclusion of personally significant material.

The debt of projective theory is not confined to the term, nor to the description of mechanism. It was Freud who first made systematic inquiry



into hidden motivations and into the genetic determinants of mental life, and it is exactly these that the projective methods seek to uncover. The psychoanalytic methods for interpreting behavior, both verbal and motor, in terms of their symbolic rather than their obvious meaning; Freud's emphasis upon the unconscious; and the distinctions he drew between latent and manifest dream content, have had a profound influence upon the significance attached to projective productions. White, writing of Freud's method of dream interpretation, describes the dream as "a natural projective method capable of revealing much valuable information if only the signs can be directly read" (23, p. 218). Play techniques, although used by analysts only as secondary tools, had their origin in orthodox child psychoanalysis (75, 82, 99). Play methods today are variously used and interpreted, but many analysts still claim them as their own special prerogative (79).

In our times, the indebtedness to psychoanalysis is becoming somewhat more even. Psychiatrists, including analysts and psychoanalytically oriented psychologists, are going beyond the traditional techniques of free association and dream interpretation, and are turning with increasing interest to the illuminating and time-saving data which the projective methods appear able to provide (155, 156).

2. "*Global*" theory. The period in which projective methods have developed is pervaded by revolt against what has been termed the *atomistic* tradition of the early experimental psychology, especially behaviorism, represented in the personality field today by investigations concerned with trait lists, rating scales, psychographs and other *objective* methods. Tests such as the Bernreuter and other personality inventories purporting to measure such traits as introversion-extraversion, dominance-submission, neuroticism and the like, imply what Allport has called *omnibus* or *sum-total* definitions of personality (210, p. 43).

*Atomistic* research is alleged to begin with the attempt to analyze psychological phenomena into elements. Opposed to this viewpoint is one which has been variously termed *global*, *holistic*, *organismic*, or *field theoretical*. Lewin's topological concepts (249), Allport's version of William Stern's *personalistic psychology* (210), Murray's adaptation of organismic theory (14), and the dynamic approach recently advocated by Maslow (253), differ somewhat in conceptualization, but unite in stressing the importance of totality and wholeness. Murray, whose theory of personality leans heavily upon the organismic biological viewpoint, quotes the following from E. S. Russell:

The organism is from the beginning a whole, from which the parts are derived by self-differentiation. The whole and its parts are mutually related; the whole being as essential to an understanding of the parts as the parts are to an understanding of the whole (14, pp. 38-39).

Murray and Maslow have both emphasized a rather fundamental division among psychologists, not only in the holistic-analytic controversy, but over the entire range of theory construction. Maslow, who describes his view as "holistic rather than atomistic, functional rather than taxonomic, dynamic rather than static, dynamic rather than causal, purposive rather than simple mechanical," points out that writers who "think dynamically" are also inclined to think "holistically rather than atomistically, purposively rather than mechanically, and so on" (253, p. 520).

The emphasis which Maslow places upon the word *dynamic* demands a brief digression at this point, since the word has been widely and often indiscriminately used, and hence has been subject to criticism. The term is, apparently, as essential in the vocabulary of the psychologist who regards interactions between parts of a system as more important than the parts themselves, and who needs to describe the process of complex change *per se*, rather than a succession of static frames run together in a moving picture. To the *dynamic* psychologist, the *moving* picture cannot be described in terms of the sum of cross-sectional views. The accusation of vagueness is, perhaps, inevitable since the term refers to phenomena which themselves lack precision. Fairly representative of current usage is the following definition:

Since psychology deals only with motion—processes occurring in time—none of its proper formulations can be static. They all must be dynamic in the larger meaning of this term. Within recent years, *dynamic* has come to be used in a special sense: to designate a psychology which accepts as prevailingly fundamental the goal directed (adaptive) character of behavior, and attempts to discover and formulate the internal as well as the external factors which determine it (14, p. 36).

Murray, like Maslow, attempts to brand psychological sheep and goats by setting the *peripheralists* off against the *centralists*. Peripheralists, he states, are attracted to observable things and quantities; they prefer to confine themselves to measurable facts. For them the data of psychology are environmental objects and physiologically responding organisms: bodily movements, verbal successions, physiological changes.

If the peripheralists ever do indulge in speculation about what goes on within the brain, they usually fall back upon the conceptual scheme which has been found efficient in dealing with simpler partial functions. . . . Men of this stamp who study people usually come out with a list of common action patterns and expressive movements, though occasionally they go further and include social traits and interests (14, p. 7).

The centralists, on the contrary, are attracted to subjective facts, such as feelings, desires, and intentions. Their terminology is subjectively derived and "they do not hesitate to use such terms as wishes, emotions, and ideas" (14, pp. 8 ff.). They are conceptualists rather than positivists.



It may be that we encounter here an example of what Seashore has called the *all-or-none fallacy* in the acceptance or rejection of a given viewpoint (263, p. 605). On the other hand, these differences reflect an age-old conflict between empiricism and rationalism, positivism and phenomenology, which has been felt in the physical sciences as well as in psychology. Whether or not this is a pseudo-issue or a real one, the hope for early reconciliation is not bright, as long as the *globalist* yawns in boredom over the statistics of the *specifist*, and the latter sneers at the constructs of the former. We should, perhaps, cultivate tolerance toward the extremists in both camps and recognize the stimulus value of controversy. We can also admit the truth of Murray's comment that "personology is still in diapers, enjoying random movements" (14, p. 6), and hope that as the infant matures his activity will become better channelized.

The special relevance of the global, conceptual approach to projective methods is clarified by an example which Maslow has used, stressing the point that his attack is not upon science itself but upon *one* possible view of science, which he calls the *reductive effort*, i.e., the attempt to analyze psychological phenomena into fundamental variables without taking account of unity and interaction.

If we take an example, such as blushing or trembling or stammering, it is easy to see that we may study this behavior in two different ways. On the one hand we may study it as if it were an isolated, discrete phenomenon, self-contained and understandable in itself. On the other, we may study it as one expression of the whole organism, attempting to understand it in its richness of inter-relationships with the organism and with other expressions of the organism. This distinction can be made clearer if we make the analogy with the two possible ways of studying an organ like the stomach. (1) It can be cut out of the cadaver and laid on the dissecting table, or (2) it can be studied *in situ* in the living functioning organism (253, p. 516).

Projective methods, it is claimed, are one means by which aspects of personality may be studied without distortion. Harrison suggests that "the global approach at least respects the complexity of personality problems and seeks some elementary understanding before bursting into figures" (154, p. 50).

Before leaving the discussion of holistic viewpoints, brief notice should be taken of the fact that, at least in psychology, this approach had its inception in *Gestalt* theory, which began with Wertheimer's well-known studies in perception about 1910. Aside from its indirect theoretical influence, the *Gestalt* experiments in the patterning of perceptual experience have a more direct bearing on projective techniques. If personality is defined as "a dynamic process of organizing experience" (5), the manner in which a person perceives is quite as important as how he behaves or what he says. This has

led to great interest in the selective response of individuals to ink blots or nonsense sounds, and in the formal structures imposed upon expressive mediums. Tests such as the Rorschach (102-142), Murray's Thematic Apperception Test (TAT, 143-173), or Shakow's Tautophone (174-178), depend on selective attention and perceptual organization; hence the determinants of response can be studied as well as the response itself. Concern with these determinants and with the way in which they are patterned by subjects is reflected in the formal scoring categories of the Rorschach, in certain approaches to TAT analysis (145, 160, 173), in Benders' analysis of form qualities in children's games and drawings (29, 64), in Kerr's and Wertham's analyses of form elements in the Lowenfeld mosaic test (193, 204), and in Erikson's use of spatial variables in describing the play of children (72, 73, 74). It should be emphasized that, although the formal approach to scoring is no more exclusively dependent upon specific *Gestalt* doctrine than content analysis is upon orthodox Freudian dogma, a close relationship exists.

3. *Support from general science.* L. K. Frank, approaching the problems of personality from the standpoint of a student of psychology as one among other sciences, finds much to support the newer concepts in personality research and the methods based upon them. His important article, quoted at the beginning of this review, ought to be read in its entirety for its abundant illustration from fields in which the busy psychological specialist has little time to become oriented.

With reference to the controversy discussed in the preceding section, Frank shows that the older sciences, including physics, have been forced to develop new approaches demanded by a changing view of the space-time universe, but have been able to digest revisions of theory with less heartburn than psychology has experienced.

Theoretical physics has adjusted itself to the conception of a universe that has statistical regularity and order, and individual disorder, in which the laws of aggregates are not observable in the activity of the individual making up these aggregates. Thus, quantum physics and statistical mechanics and many other similar contrasts are accepted without anxiety about scientific respectability. The discrete individual event can be and is regarded as an individual to whom direct methods and measurements have only a limited applicability. We can therefore acknowledge an interest in the individual as a scientific problem and find some sanction for such an interest (5, p. 395).

Frank compares our factor analyzing and trait isolating techniques to the analytic-destructive methods in nineteenth-century physics, which required the breakdown of the substance studied, and hence failed to reveal their true nature. For example, physical phenomena such as temperature and light, which were once studied in isolation, are now seen as transformations of energy variously manifested. In the indirect approaches to person-



ality through a study of its projection in *neutral situations* or *unstructured fields*, Frank finds a parallel to the use of electric current or polarized light in determining the composition of substances through their effect upon these media. It is just such methods as these that have led modern physics to the discussion of processes *within the atom* and to an interest in the behavior of individual electric particles which are themselves not directly observable. Similarly, he holds that personality, which is not observable in essence, can be understood as an organizing process through its projection on the screen of a meaningless ink blot or a formless chunk of clay. If these criteria appear subjective and incredible, Frank points out that certain methods in physics are also open to criticism.

Personality studies by projective methods have not, of course, been as extensively used, nor have the patterns used by subjects been so well explored. The important point is that the way is open to the development of something similar to spectroscopic and diffraction methods (5, p. 406).

Better known in general psychology, and more influential outside the field of personality, is the physicist Bridgman, whose introduction of the "operational definition" has had a far-reaching influence on psychological thought (see Kantor's discussion, 244). For Bridgman, a concept must be defined in terms of the operations by which it was derived; for example, dimensions are defined in terms of meter sticks and time by clock readings (220). Frank's definition of personality as process, "a way of living and feeling," or "a manner of organizing and patterning the life situation" leads to an operational definition of personality as that which an individual does in situations described as projective.

If it appears that the subject projects similar patterns or configurations upon widely different materials and reveals in his life history the sequence of experiences which make those projections psychologically meaningful for his personality, then the procedures may be judged sufficiently valid to warrant further experimentation and refinement. In undertaking such exploration, the experimenter and clinician may find reassurance in the realization that they are utilizing concepts and methods that are receiving recognition and approval in scientific work that is today proving most fruitful (5, p. 409).

#### HISTORY OF RELATED RESEARCH

1. *Studies in imagination and phantasy*. In 1930, when the writer happened to undertake a paper on the subject of phantasy, very little experimental work had been done from the individual point of view. Empirical material was available in psychoanalytic case studies, but emphasis was placed upon the description of typical, *universal phantasies*, such as the *Oedipus phantasy* (247), the *phantasy of illegitimacy* (248), *phantasies of*

*rebirth* (252), and a number of others (228, 230, 237, 259). Much space was devoted to the symbolism of phantasy and its relation to dreams and myths (230, 252), but although it was recognized that the personal meaning of daydreams is less disguised than in dream content, and that the subject is more loath to disclose them (247), interest in techniques for eliciting such material had not developed. L. P. Clark, apparently, must be credited with the first attempt to stimulate phantasy, using it as an approach to narcissistic patients incapable of transfer (224).

The psychological literature, although it reflected the growing emphasis upon affective and conflict-solving aspects of imagination, was concerned with phantasy chiefly as one manifestation of thought; as a psychological process rather than as a key to individual mental life. Varendonck had produced an extended analysis of daydreams, based on an introspective examination of his own foreconscious activity (269), and a systematic developmental study of daydreaming had been published by Green (234). Lehrman had written several articles on the compensatory nature of normal and neurotic phantasy (247, 248); imaginary playmates had been studied as the projection of young children's needs and wishes (237); and Conklin, in one of the earliest attempts to test psychoanalytic theory by a psychological technique, had conducted a questionnaire study of the *foster child phantasy* as recalled by college students (225). Although the Rorschach test was nine years old, already in use in Europe, and arousing the interest of a small group of American workers, there were few references to it in the English literature.

If the individual approach had been neglected, an experimental attack on imagination as a mental function had not. Galton, among his many interests, began investigations of imagery in 1883 (231). Binet and Simon used ink blots in early tests (217), and this use was extended by American experimentalists including Whipple, Dearborn, and G. S. Hall. (See review of early work in Krugman, 128.) Although these studies were largely concerned with cognitive rather than emotional aspects of phantasy, they form one branch of the ancestry of present projective methods because of the materials and techniques suggested. The most recent experimental approach to developmental aspects of imagination and its relationship to cognitive thinking was published by M. D. Vernon in 1940 (260).

The earliest direct forebears of the best known and most widely used of modern projective methods (Rorschach and Thematic apperception) were Britain's investigation in 1907 of imagination by means of compositions written in response to pictures (221), and the use of ink blots by Bartlett, another English researcher, in 1916 (214). The latter, in his use of the blots, went beyond mere investigation of imagination, and speculated on differ-



ences in intelligence, background, vocational interests, and the like. Content was analyzed, and in many ways the handling of the materials bears striking similarity to the method which Rorschach later developed.

Hermann Rorschach, Swiss psychiatrist, first published his *Psychodiagnostik* in German in 1922 (135). As noted above, Rorschach was not the first to use ink blots, but he was the first to develop a workable method (a *shorthand* as Beck has called it) for handling the complex individual response pattern. Rorschach himself died shortly after the publication of his famous work, but a tremendous amount of research has followed. The test is widely used in clinical diagnosis (103, 104, 108, 124), has been introduced in the armed services for research and diagnostic purposes (112, 113), and has been applied to a variety of problems, including psychopathology (104, 124, 133), developmental psychology (122, 126, 127) and recently to vocational guidance (134), not to mention innumerable other problems of personality research. (For comprehensive reviews and bibliographies, see references 116 and 128.)

Historical milestones in the progress of Rorschach research in this country have been the publication in 1924 of the Rorschach-Oberholzer monograph in the *Journal of Nervous and Mental Diseases* (136); Vernon's article in 1933 which called attention to the success of the test in Europe (138); the publication of Beck's *Introduction to the Rorschach Method* (103) in 1937 (the first systematic guide to administration, scoring, and interpretation); and the founding of the Rorschach Institute in 1939, with Klopfer as its guiding spirit and the *Rorschach Research Exchange* (established in 1936) as its medium of communication. In 1942, two books appeared almost simultaneously, one by Klopfer and Kelly (125), and a somewhat overpopularized presentation by Bochner and Halpern (108). At the time of this writing a new book by Beck is in press (106). There is probably no topic in psychology concerning which pro and con feeling runs higher, but despite the rash enthusiasm of converts and the blind opposition of skeptics, it has grown in use and reputation. White refers to the Rorschach as "a good example of that happy combination of genius and hard work which the study of personality so sorely needs" (23, p. 227).

Another important family of projective tests made its appearance in the American literature in 1935 with the publication of Morgan and Murray's paper on a method for investigating phantasy (157). Here the basic technique of the Thematic Apperception Test (TAT) was first described. The test consists of pictures which the subject is asked to use as illustrations for plots of his own creation. Certain investigators, Harrison among them (152), still prefer this earlier series of pictures to later modifications introduced in the set now provided by the Harvard Psychological Clinic (158).

Historically, as well as from the point of view of standardization, these two methods (the ink blot and the picture-story) are the aristocrats of a growing clan. Others will be mentioned later; at present we are concerned with beginnings and major trends. Among these was a five-year investigation of children's phantasy published by Griffith in England in 1935 (235). This study, which utilized imagery tests, ink blots, dreams and drawings of fifty normal children, led to the conclusion that phantasy is one of the ways in which children deal with their problems, and hence should be viewed not merely as withdrawal from reality but as an aspect of adjustment. This finding is important both for diagnosis and for therapy.

2. *Studies in word association.* It is a well known historical fact that interest in "the association of ideas in the mind" considerably antedates the beginnings of experimental psychology. The preoccupation of philosophers with this topic gave to the British Associationist school its name. The history of the movement leads from the speculations of John Locke down through the conditioning experiments of the modern behaviorists. Galton's work in 1879, Wundt's in 1880, and the introduction of association experiments into American laboratories by Cattell, Munsterberg, and Jastrow before 1890, marked the beginning of scientific interest in associational activity (226).

As in the studies of imagination, interest from the individual standpoint developed first in psychiatry, followed a parallel court, and finally gained enough momentum for recognition by psychology in general. It was not until the publication of Jung's first studies in word association in 1904, followed by the extensive experiments of Kent and Rosanoff in 1910 that the significance of association tests for personality study became impressive. Moreover, in the statements of early users of the method, we find concern with a problem which we have already seen as central in modern personality study: that is, the search for techniques which not only add to our knowledge of what Titchener called the "generalized, normal, adult, human mind," but also serve as an approach to the whole, unique personality. Bleuler, in his chapter in Eder's translation of *Studies in Word Association*, writes:

In the activity of association there is mirrored the whole psychical essence of the past and of the present with all their experiences and desires. It thus becomes an index of all the psychical processes which we have but to decipher in order to understand the complete man (218, p. 4).

Eder points out in his own introduction to Jung's *Studies* that the departure we owe to Jung is the application of the association method to unconscious mental processes, and the theory of unconscious complexes (243). Thus it represents the first effort to study the deeper strata of personality by an experimental technique. Again stressing the importance of studying psychic functions in context, Jung writes:



We must bear in mind that the association experiment cannot deal with a separated psychic function, for any psychic occurrence is never a thing in itself but is always the resultant of the entire psychological past (242, p. 225).

The work of Kent and Rosanoff which led to the compilation of frequency tables for the evaluation of common and unusual responses to word lists, and their efforts in the direction of standardizing both technique and interpretation (245, 260) were followed by numerous other studies by Hull and Lugoff (238), Woodrow and Lowell (274), and Woodworth and Wells (273), to cite only a few. An excellent bibliography of this material is furnished in H. R. Crossland's monograph published in 1929 (226). Of special interest in conjunction with projective methods is the fact that similar problems of standardization were encountered. Wells and Woodworth point out that standardization cannot easily be accomplished. They conclude, however, that such difficulties do not detract from the significance of association techniques.

Few procedures in experimental psychology have so richly rewarded their investigators with the possibilities of practical application as the association method. . . . Within the past seven years it has achieved and bids fair to hold indefinitely its place in the foremost rank among the methods of individual psychology (273, p. 73).

Although the method has been supplemented to some extent by more flexible techniques for tapping affective thought and phantasy, this estimate does not appear extreme even today.

3. *Studies in language.* Although the recent revival of interest in language analysis has a history of its own and is not strictly a branch of projective research, the two topics have several points in common. The first of these is an interest in the formal qualities of expression.

Language traditionally has been known as the "vehicle of thought" with the thought attracting far more attention than the vehicle. But there are those who object to the traditional distribution of attention on the ground that the vehicle as well as the freight should be given systematic scrutiny (261, p. 181).

Piaget's extensive investigations of children's language usage (257, 258) as an approach to the ontogenetic evolution of thought and socialization, have stimulated a great deal of interest in this area. Although his emphasis is upon establishing universal generalizations appropriate with reference to the progressive developmental stages of social and moral consciousness, rather than the clinical study of individuals, Piaget's recognition of the manner in which language forms (such as egocentric expressions and causal statements) may reflect the *inner* or emotional level of the speaker, bears a significant relationship to the theory underlying projective experimentation.

Southard, the psychiatrist, made an earlier application of formal language analysis to personality structure, pointing out a similarity between the use of the four grammatical moods (imperative, indicative, subjunctive, and optative) and the traditional temperaments: choleric, phlegmatic, melancholic, and sanguine (265). Grings has recently used grammatical classifications in analyzing responses to *Tautaphone* records (174), and Balken and Masserman have suggested others which have been applied in studying the TAT protocols of neurotic patients (145). Among the categories utilized, these authors revived Busemann's verb-adjective ratio and offered further evidence that high quotients are associated with anxiety and instability (144, 145). Johnson (239, 240) and Boder (261) have also used this and other formal counts in studies of written materials.

Research in language undertaken from this viewpoint is compatible with Frank's conception of personality structure. If a personality does, as he assumes, organize experience in terms of its "own private idiosyncratic world of meanings," it is logical to extend this assumption to include an individual's choice of language for self-expression as one aspect of personality. If this line of research is to prove truly productive, results are likely to come, as Sanford suggests, from a search for psychologically meaningful categories, in place of strict adherence to traditional grammatical modes of analysis (261, p. 831).

The recent interest in semantics is another expression of the conviction that no manifestation of personality is meaningless or outside the legitimate field of investigation (255, 272). Although the semanticists appear to insist on a peculiar reversal of what would seem the more acceptable conclusion, in assuming that bad semantic practices produce rather than result from maladjustment, their work has focused attention on possible relations between speech and personality (241, 246). Since many of the projective methods elicit a verbal response expressed in a subject's own idiom, analysis of language forms has direct bearing on these problems.

4. *The use of "personal documents."* Gordon Allport has long been interested in what he terms *idiographic* as opposed to *nomothetic* research. The former is simply another name for research concerned with the single case, and the latter is used to describe normative science which seeks to establish uniformities. In the personal document Allport sees the ideal datum for the intensive study of individuals. His recent monograph (209) traces historically the use of such materials; discusses the possibility of quantitative treatment from the idiographic point of view; points to various studies which offer methods for judging reliability and validity, and cites a number of examples showing their usefulness. Since Allport's definition is applicable to any projective protocol, it is pertinent here:



The personal document may be defined as any self-revealing record that intentionally or unintentionally yields information regarding the structure, dynamics and functioning of the author's mental life (209, p. xii).

In defending the use of such data (which until recently have been considered worthless for scientific purposes because of their subjective nature) Allport discusses Clifford Beers's classic *A Mind That Found Itself* (216) and comments:

From the point of view of controls few documents are worse, yet from a pragmatic point of view we are warned that scientific safeguards will not in themselves save a poor document from the dust pile nor prevent a good one from contributing to the course of scientific progress (209, p. 11).

We are probably justified in the following inquiry: Suppose we study carefully Beers's document. Suppose, also, that this source of information is supplemented by other data which enable us to know the writer so well that we might predict his subsequent acts, and might even conceivably express relationships within his personality in numerical terms, as Baldwin did in a single case study using intra-individual statistics (212). What have we added to the science of personality? What does our knowledge of Beers contribute to our knowledge of others?

Allport would reply that this question is raised from the nomothetic point of view and is the result of thinking only in terms of generalization and comparison. His emphasis is expressly not upon the study of an individual case as a basis for generalization but, rather, as a legitimate scientific end in itself. Although Frank has rallied support from other disciplines for single case research and Lewin has argued its scientific respectability (249), on the basis of Allport's presentation it seems desirable to go beyond the rationalization "They do it, why can't I?" which imitators of physical science are prone to use in support of borrowed methods. Allport does not explicitly state the following answer, but it runs throughout his monograph by implication. It is necessary to assume that the final justification for the study of particular individuals lies neither in what we learn about personality *in general*, nor in what we learn about Beers, or any other unique individual. It is, rather, what we learn about *how* to know Beers that provides an approach to other case studies, thus leading in the direction of valid prediction. The predictions for which validity is sought are not predictions for other individuals on the basis of one, but for one individual based on sample observations designed to illuminate the dynamics of his own personality.

Although Frank and others grant that investigations based on such a philosophy are in the exploratory stage, preoccupation with the problem of controlling observation technique in such a way that the behavior studied is not distorted nor its unifying principles obscured distinguishes the newer

approach. The special questions which arise in relation to control of experimental conditions are discussed later under *Techniques* and in the section on *Methodology*. It is sufficient to note here that single-case research, the use of personal documents, and the projective techniques all share in common the objective of discovering sound and reliable, and at the same time more penetrating and comprehensive methods for the study of personality.

## *Applications*

There are several ways in which projective methods may be classified. They may be grouped according to:

1. the nature of the materials used for projection,
2. the functional use which the subject makes of the materials,
3. the techniques of presentation used by the experimenter, and
4. the purposes which govern the various applications.

The diversity of the methods themselves, and the variety of ends served can best be illustrated by considering the techniques under each of the above headings.

*Materials.* Ink blots, as we have seen, are the oldest and from many points of view the most versatile of the materials used. In the first place, they have a definite structure which makes it possible to classify responses in terms of specific determinants, an advantage which is lacking in more vague mediums, such as Stern's cloud pictures (202). At the same time, they are less meaningful in terms of personal experience than pictures, which may have rather definite associations. An additional advantage of the blots is that they elicit a response in which factors can be scored both for content and for the characteristics of its formal pattern.

Pictures, such as the TAT set (158), are about equal in popularity, and although they are subject to the limitations discussed above, they often provide a more readily intelligible sample of thought content for interpretation. Amen developed a series of pictures for use with children which included silhouettes and movable figures (143). Symonds has set up criteria for the selection of pictures suitable for use with adolescents (170). In the Szondi test, pictures are presented in pairs, the subject being asked to choose the one preferred or disliked (2, 200). Horowicz has made use of a similar technique with children (189, 190, 191).

Other materials include stories which the subject is asked to retell or complete (159, 183, 186, 187). Art media, such as clay modeling, finger painting, and drawing have been extensively used (25-47). Materials such as blocks



(57, 58) and mosaic patterns (193, 204) have also proved effective. Dramatic play, ranging from the puppet shows used by Bender, Woltmann, and others in group treatment (48, 52, 53, 56) to the psychodrama introduced by Moreno and his coworkers (49, 50, 54, 55) has been utilized. A unique device introduced by Skinner and adapted by Shakow, Grings, and Trussell, is the so-called *Tautaphone* or verbal summator (174-178). This instrument produces low vowel sounds which resemble speech. Subjects are asked to tell what the voice is saying, thus projecting their own preoccupations and meanings into an auditory medium. Pickford made use of nonsense syllables as projective materials (198) and Mira turned the simple motor task of drawing lines into a *myokinetic* technique for studying various aspects of personality (196). A recent attempt has been made by Sargent (201) to incorporate projective principles in a "paper and pencil" personality test for group use.

*Functional uses.* Frank (5) distinguishes:

1. *constitutive* methods, in which the subject imposes structure upon unstructured or partially structured materials;
2. *interpretive* methods, in which the subject is led to express or describe what a stimulus situation means to him;
3. *cathartic* methods used for the discharge of affect; and
4. *constructive* methods which call upon subjects to organize and arrange materials according to their own conceptions.

There is, of course, considerable overlap in this classification. Constitutive methods are exemplified by the various plastic media. The one criterion is that the material should be amorphous and easily adaptable for subjective expression. The interpretive-methods category is illustrated by ink blots and pictures. The attribution of meaning is not, however, confined to such stimuli. Whether the subject molds his own object or responds to one ready made, subjective interpretations direct or symbolic, verbalized or implied, are usually evident. Likewise, discharge of affect is thought to be present to some extent in all of the experimental situations. It is generally assumed that the subject sees in an ink blot what he *must* see in order to solve unconscious conflicts; in his story themes he voices his dilemmas and arrives at emotionally satisfying outcomes.

Ordinarily, the cathartic function is most evident in techniques which are suited to the release of aggressive impulses, as in certain play techniques. Despert's experiment involving the use of a knife (182) is an interesting attempt to provide both for the expression of aggression and for its eventual reorganization into more socialized activity: Children were allowed to use a knife to scrape cardboard into shreds (destruction); next they were given water with which to mix the materials into a plastic (gratification of the sup-

posed impulse to *mess*); and finally they were encouraged to use this product for the construction of masks for use in play.

Finally, under constructive methods we may list blocks, mosaics, and building materials, as well as such highly specific objects as dolls, household articles, and *world toys* (181, 194) from which the subject constructs his *microcosm* (a term used by Erikson [74] to describe the miniature life scene depicted in the play model). Wertham has developed differential diagnostic criteria for the mosaic test, which he advocates as simpler and less time consuming than the Rorschach. Patterns are derived which in many respects parallel the Rorschach findings for various personality syndromes found in psychoses, neuroses, and organic conditions.

*Techniques.* Classification by techniques may be made both on the basis of differences in experimental control and in approach to interpretation. Control is introduced first according to the nature of the chosen stimulus. Degree of structure varies on a continuum from unstructured clay to definitely representative toys. Here a paradox is introduced. The more structured the material, the more limited is the subject's range of response. Hence, *subject* controls vary inversely with control over the essential *neutrality* of the situation itself, thus defeating the very purpose of control, which is to provide a standard by which response variations may be judged. As structure is increased, meaning also increases and there is less opportunity for personally significant differences from subject to subject to appear. Moreover, when differences do appear, the impossibility of estimating effects of experiential background furnishes an uncertain ground for comparison. The same is true when special techniques of presentation, such as paired comparisons, set questions and the like are introduced, since there is no way of equating their stimulus value for different individuals.

This has been called the *stimulus fallacy* by L. L. Thurstone. Frank, in his introduction to the Lerner-Murphy monograph (7), cites an early paper of Thurstone's in which he expresses the view that a subject creates his own stimulus to which he then responds in characteristic manner, hence a stimulus situation, no matter how well *controlled* does not always mean what the tester intends (268).

In projective experiments the fallacy is turned to account and the variegated qualities of a stimulus not only cease to be a disadvantage, but become of crucial importance. For example, when a subject sees sexual symbolism in an ink blot which others view as a landscape or as an animal form, or when Bender's children spontaneously produce such material in drawings (28), it is more logical to regard these productions as individually significant than the attitudes elicited by Levy's controlled and highly suggestive procedure. The latter presents the child with dolls and clay, the child is encouraged to make breasts for the mother, and the baby is put to nurse. This method is highly successful in drawing out sexual attitudes as well as hostility, but how much of this effect should be attributed to the technique, and how much to the child is a debatable question (83).



The range of procedural variations is illustrated by contrast between the standardized methods of Levy (83, 84), Ackerman (57, 58), the somewhat more flexible techniques of Conn (66, 67, 68), and Solomon (95, 96), and the almost complete freedom of action maintained in some free play situations (74).

Interpretive systems for handling responses range from the empirical analyses of Harrison (152, 153) to conceptualization based on psychoanalysis or the Murray press-need system (14). Interpretations also vary in the relative emphasis on form and content. In Bender and Woltmann's reports of group responses to puppet shows (48, 56) and Despert's analysis of play situations (69, 70, 71, 182), symbolic interpretations of content are emphasized. In the mosaic test (193, 204) form is paramount. In Homburger Erikson's studies, content is interpreted from the psychoanalytic viewpoint but spatial variables are also introduced and interpreted in terms of configuration (72-74). Of the two most extensively used tests, the Rorschach and the TAT, the former relatively stresses form and the latter relatively emphasizes content. These two, however, fall somewhere near the middle range in the use of both scoring approaches.

*Purposes.* An examination of projective methods with regard to the ends they serve shows three major purposes for which they are commonly used. These may be subsumed under the topics of diagnosis, therapy, and experiment.

1. *Diagnosis.* As already indicated, projective methods developed first in the clinical setting as devices for tapping phantasies, sampling conscious and unconscious thought, and revealing the characteristics of individual perceptual organization. As a psychiatrist, Rorschach's primary aim was to provide a method which would help to illuminate dynamic factors in mental disease and serve as an aid to differential diagnosis. Recent reports on the clinical use of the test, and the demonstrations of fairly stable response patterns for various pathognomic conditions, have gone far toward the realization of that aim (103, 105, 125). The more recent TAT is also demonstrating similar potentialities (151, 160, 161).

In child analysis the special diagnostic value of play, which has been described as the child's language, a medium of expression which precedes facility in verbal communication, was first recognized by Melanie Klein. Anna Freud has also been a pioneer in the use of play techniques, but she has evolved an approach which is far more conservative than Klein's. Freud (75) uses play primarily to gain rapport and insight, reserving *deep* analysis for later stages of therapy; Klein, however, regaled her fascinated young patients with symbol interpretation from the beginning of the first contact (82). Following these leads, play situations were used extensively by others as an ad-

junct to both diagnosis and therapy (57-100). Sargent (93) has reported an observation of a *normal* child which tends to confirm observations and interpretations of play made under clinical conditions. As a diagnostic method play has been used more informally than formally, but certain controlled techniques are being developed (67, 84) and other possibilities, such as the adaptation of play methods as a *readiness* test for preschool children, have been explored (97). Mayer and Mayer (87), Rosenzweig and Shakow (91), and Murray (14) have adapted certain play methods for use with adults.

2. *Therapy*. It was soon discovered that expressive methods, especially play, art techniques, and drama, frequently served a double purpose. Insight into pathogenic factors and adjustment mechanisms was obtained, while at the same time the release which the projective situation both stimulated and permitted, appeared therapeutic in its effect. The literature is replete with case material indicating that play therapy can bring about reorganization and reintegration of attitudes (65, 81, 85, 98). Some workers attribute the benefit to the release and desensitization provided by play itself (83, 84, 88); others believe that improvement comes indirectly as the result of relationship with the therapist which the latter uses in treatment (59, 65, 75, 79); still others hold that insights developed in play and verbalized by the child are essential for therapy to take place (66, 67). Some form of interpretation, support or suggestion from the therapist, directly or indirectly given to the child, is considered necessary by Cameron (65), Liss (85), Solomon (95, 96), and others but all except Klein (82) recommend extreme caution in its use.

Whatever the explanatory theory, experimenters with projective techniques are in substantial agreement that emotional growth and symptom remission can and often do result, even when the methods are used primarily for diagnostic purposes. Experience has demonstrated, in fact, that a rigid distinction between diagnostic and therapeutic contacts is unjustified and perhaps dangerous (97). In the course of therapy, information is gained which leads to greater precision in diagnosis; whereas even in the short diagnostic session the subject undergoes an experience which can be traumatic, or therapeutic, or neither, depending upon the emotional state of the patient, and the handling of the relationship with the clinician. Even in a relatively stereotyped situation, such as an intelligence test, favorable or unfavorable changes may be initiated, as in the case of a child seen at the Psychological Clinic at Northwestern University who, following an intelligence test, announced to his mother that the examiner had "taught him to concentrate." In the projective situation such changes can be quite radical, depending upon how effective that situation is in releasing impulses, and upon the manner in which the resulting anxiety is handled. These facts have caused many writers to urge that only trained persons should use projective



methods, and that experienced psychologists should handle them with caution (78, 79). Certainly the practice of interpreting the results of projective tests to subjects and, for that matter, the irresponsible passing out of test scores of any kind, cannot be too severely condemned.

3. *Experiment.* Extensive use of projective methods for research purposes has been made in the Harvard Psychological Clinic under the direction of H. A. Murray. An ambitious personality study of fifty-one paid subjects (most of them Harvard students) utilized projective tests such as the Rorschach, Murray's own TAT, and certain tests devised specifically for this investigation. Among the special tests were *imaginal productivity* tests (including story completion, *similes*, and *Beta* ink blots), *dramatic productions*, and *musical reveries*. The battery also included a variety of non-projective tests including measures of hypnotizability, level of aspiration, reactions to frustration, and the like. Conspicuously absent were the usual inventories and ratings. Commenting on the omission, Murray launches another attack:

We included some of the procedures commonly employed in academic studies; intelligence tests and a variety of questionnaires, but these contributed little to our understanding. American personologists base their conclusions on a much larger number of subjects than we studied, and in this respect their findings are more representative than ours. What they usually study, however, are the physical attributes of movement, manifest traits and superficial attitudes, facts which subjects are entirely conscious of and quite willing to admit. Thus their researches do not penetrate below the level of what is evident to the ordinary layman. . . . The original data, then, are of uncertain value, and no amount of factor analysis can make them more reliable (14, p. 33).

At Sarah Lawrence College, Lerner and Murphy and a group of collaborators have conducted a four-year study of nursery-school children in which projective techniques have played a leading role. Their equipment included *miniature life toys*, *sensory toys*, and plastic materials such as dough and cold cream. Group play formed a part of the experiment, and several ingenious *active play techniques* were introduced (7).

Both the Murray and the Lerner-Murphy experiments were exploratory in nature; designed to gather a variety of information "which would somehow illuminate personality development." Lerner and Murphy state that they have "accepted the general focus of working toward a small collection of intensive case studies" (7, p. viii). Their announced aim is to gather cross-sectional data at a later time when the intensive studies have given direction to the search.

In other experiments, projective data have been used to throw light on many different psychological problems. Horowicz used paired pictures in developmental studies of race prejudice (190), self-identification (189), and social conformity (191). Aggression in children has been the subject of sev-

eral studies, utilizing various techniques; Baruch used a free-play technique with young children (61); Fite has made use of rather specific questions and comments to stimulate aggression in her subjects (185); and destructive behavior has been investigated by Ackerman in an ingenious experimental set-up using blocks. In the latter study, the child upon entering the room was confronted with two rows of toy structures, one torn down ready for building, the other set up to invite demolition, thus offering an opportunity to study the contrasting behaviors of different children (57, 58).

Frustration studies have utilized projective techniques as auxiliary tests (162, 167). Bellak investigated the projection of aggressive feelings in TAT stories told after the first five plots offered by a subject had been severely criticized (148). Sarason has applied the TAT in experiments with feeble-minded girls (166), and Slutz discussed its value for developmental studies (169). Investigations of attitudes using projective devices have been carried out by Tuddenheim (203), Dubin (184), and Proshansky (199). Tuddenheim used a *reputation test* as a measure of projected attitudes; Dubin predicted the answers to attitude tests on war, government and labor on the basis of play constructions built by undergraduates on the subject "the world as I see it." Proshansky's experiment involved comparisons between replies to a labor attitude scale and interpretations of pictures which had been previously judged ambiguous as to "outcome for labor." Adaptations of other techniques have been turned to projective use, as in Buhler's study of neurotic and normal performance on the 1916 Stanford Binet *ball and field* test (180), and in Marquidt's use of eight simple oral questions based on an idea suggested by Binet (195). All of these studies present some degree of positive results in the direction expected according to the hypothesis of projection.

### *Experimental Results*

Projective methods are not yet ready to fill a bulky section on *experimental results* if by that title is meant a list of findings in crucial experiments. The preceding section gives a cursory view of the problems to which they have been applied, and indicates the general trend of results. Although a great deal of incidental information has been contributed to personality study, psychopathology, and special problems, it seems important at this state of development to examine results which support theories on which the techniques are based, and the findings of experiments designed to test the efficacy of the methods, rather than to classify miscellaneous outcomes of innumerable investigations covering a wide range of psychological phenom-



ena, from psychiatric problems to social attitudes. As long as projective methods remain in the exploratory period, research interest should be focussed on questions of method itself. In the final section of this survey, methodological problems will be discussed. At present it is necessary to examine a few outstanding researches which bear on underlying assumptions and predictive value.

Two experiments, one by Sears and one by Murray, are worth noting because of their attempt to demonstrate experimentally the basic mechanism of projection. Sears used a rating scale technique in which a group of fraternity men were asked to rate themselves and others on the possession of certain traits. On the assumption that some individuals would use projection to protect themselves from the acknowledgment of undesirable traits in their own personalities, Sears further assumed that the largest amount of projection would occur in those who lacked insight. A subject's average rating by others was used as the estimate of his *true* score in a trait; his average score in attribution of the trait to other individuals was taken as the measure of projection, and agreement between self and others' ratings provided the criterion of insight. Results were in the expected direction: those who were more *stingy* than average, and who lacked insight into that fact, rated others higher in stinginess, on the average, than did those who were equally stingy but recognized this characteristic. Each trait in Freud's *anal erotic* syndrome (stinginess, obstinacy, disorderliness) was given more extreme ratings by the insightful group. Sears concludes:

Without belaboring the question of whether lack of insight was a result of repression in a Freudian sense, or whether it was a conscious or unconscious process, it can be said that the effects of projection have appeared in this situation in a way that was predictable from the present redefinition of that process (19, p. 576).

Another fundamental experiment has been conducted by Murray, who has studied the effects of fear upon response to pictures (13). Two series of 15 photographs were administered to five 11-year-old children at a house party. The photographs were clipped from *Time* and the subjects were asked to rate the maliciousness of the pictured persons on a nine-point scale. Each series was administered twice, once after a *normal pleasure experience*, and once after a game of *murder*. Series A was given first under the pleasure condition preceding the *scare* game, both series were administered after the game, and series B was repeated the following morning after the effects of fear had presumably subsided. (The reality of the fear situation was further supported by the fact that one of the children actually hallucinated a burglar early the following morning!) The results show a marked increase in the *badness* scores assigned to the photographs seen under the fear condition.

A number of studies claim some degree of predictive success for projective

methods, but only two will be described in detail, one of which uses the TAT, the other a modified version of the Rorschach method.

Harrison, impressed with the fruitfulness of the TAT in clinical practice, undertook in conjunction with a qualitative study by Rotter (163), a "quantified and controlled validity study" (152, 153) on the assumption that:

Evidence for the congruence and unity of the personality would be clearly demonstrated if it were shown that an individual cannot relate what at face value are impersonal stories without revealing so much about himself that a thumbnail personality sketch including characteristic traits, biographical facts, attitudes, intelligence level, personal problems and conflicts, could with varying degrees of accuracy be written from story analysis (152, p. 123).

The test was given to 40 patients at the Worcester State Hospital about whom nothing was known to the experimenter. Personality sketches were written, and the items included were subsequently checked by an assistant against the hospital records. The index of validity used was the ratio of right guesses to right-plus-wrong, which yielded a quotient expressible as a per cent. Two controls were used. First, the themes of 15 patients were randomly matched with histories, using the same technique of checking. This furnished an empirical criterion of chance expectancy. As a second control, ten case histories were matched with ten artificial or *guessed* write-ups. Harrison obtained a validity index of 82.5 per cent correct inferences by means of his experimental technique, which was significantly higher than the indices obtained by either control method. The correlations between guessed and actual I.Q.'s was .78 with an average deviation of 9.5 points. Although diagnostic labels are themselves subject to error, guesses as to disease classification were 75 per cent correct. This figure is all the more impressive, in view of possible sources of error in the diagnosis itself as well as in the TAT interpretation.

Harrison further separated biographical from personality and intellectual items since, as the writer points out, differential validity, if it markedly favored biographical information, would detract from the practical potentialities of the method, proving the test more efficient in uncovering information easily obtainable by other means.

The comparison showed no significant difference between the two types of data, and a somewhat higher validity was found for etiological items (guesses as to motivation and causes of difficulty) than for the other item groups. In a supplementary experiment one further caution was introduced. This time the analysis was done *blind*; that is, another examiner administered the test in order to eliminate cues from direct contact with the subjects. The validity fell from 82 per cent to 74 per cent. This difference is statis-



tically insignificant and without clinical interest, since blind interpretation is a research tool which has no place in practical diagnostic work.

Rorschach validity studies utilizing the blind-interpretation technique (115, 133) as well as degree of correspondence with other criteria such as psychiatric diagnoses (107) have been promising, but because they are based on small numbers of cases and have not always been rigorously controlled, the results are interesting rather than definitive. Ruth Munroe's experiment at Sarah Lawrence College is not subject to these criticisms. Munroe's *Inspection Technique* (130, 131, 132) and Harrower Erickson's method for group presentation (113, 114) are beginning to make possible wide-scale administration which should overcome some of these objections.

For two consecutive years the individual Rorschach was given to the entire freshman class at Sarah Lawrence, the protocols were scored by the *Inspection Technique* and the results were put aside until spring when a checkup on the predictions could be made. A group of problem students selected by independent criteria such as academic failure, referral to the psychiatrist, and problem behavior observed by teachers, had an average of 7.9 deviations on the check list of problem indicators, with a range of 5.1, whereas a group of 15 selected as unusually well adjusted by teachers, had an average of 2.6 deviations with a range of 1.4. Scarcely any overlapping appears between the groups (130, p. 233). Of 43 students having five checks or less, only one appeared on any list of girls having difficulty of any sort during their freshman year, and this one girl was not, according to Munroe, considered a serious problem either by psychiatrists or teachers. Only one serious problem was altogether missed, "a girl whom two psychiatrists described as delinquent and a nuisance, but probably not deeply neurotic." The author states that errors in classification did occur. Two students who were rated D+ on a scale from A to E, should have been placed in E, the lowest group. In spite of occasional misjudgments of this sort, the test acquitted itself much better than experience has led us to expect the *paper and pencil* tests to do. The Bernreuter, administered to the same subjects in this study, completely missed some of the most seriously maladjusted students (130, 132).

### *Methodological Problems*

The variety and richness of material which the projective methods provide is at once the delight of the clinician and the despair of the experimentalist. The research worker who attempts to use any of these methods is immediately impressed both with their infinite possibilities for interpreta-

tion and insight, and the seemingly insurmountable difficulties in the way of scientific treatment of the data. Many able psychologists dismiss them as "vague"; no doubt they are justified in the choice of adjective. Others are challenged by that very vagueness, and the pioneer hope that there may be "gold in them thar hills" behind the fog. All scientific problems are unclear until ways are found to approach them.

Problems of quantification and standardization have been the chief source of controversy, not only between advocates and opponents of the methods, but among the enthusiasts themselves. To standardize or not to standardize, and if so how, is invariably an adequate stimulus for argument. It is generally recognized that the complexity of personality can no more be expressed by a psychograph of traits than it can be represented by a photograph of the physical profile. Another point on which nearly all can agree is that established mental test procedures cannot be carried over unmodified to the newer techniques.

Usual standardization procedures of mental tests have not been successfully applied to these (projective) methods, as it is the configuration of factors present rather than the independent quantity of each factor that describes the personality. Broad experience of the psychologist rather than statistically reliable norms is the necessary prerequisite for using these procedures (2, p. 76).

The misunderstandings which arise when configuration is overlooked are illustrated in a recent study of the relative efficacy of various personality tests used in a study of delinquent and normal girls (219). This experiment demonstrated that, contrary to expectation, the delinquents showed less of a certain Rorschach factor commonly associated with lack of control and impulsivity, than the normal high school group. As Rorschach workers have frequently explained, factors taken out of context have little meaning, for the reason that it is not the absolute amount of one determinant but its relation to the whole pattern which gives it significance in the individual protocol.

The application of standard methods for estimating reliability and validity further complicates the problem of quantification:

*Reliability.* In psychometrics, the usual checks on reliability have been the split-half technique, and correlations either between repetitions of the same test, or between alternate forms. None of these methods are wholly satisfactory for projective tests. For example, split-half correlations of Rorschach factors mean little because the ten cards are admittedly uneven in the type and amount of response they produce, and because the technique involves isolating factors from context. Hertz claims reasonably satisfactory results for the split-half method, but does not recommend it (116).

Repetition of projective tests is also somewhat dubious, although it has



been claimed that *basic* aspects of the Rorschach pattern are little altered upon retesting (110, 124). Fosberg found that although subjects could produce *good* or *poor* Bernreuter scores at will according to instructions, they were successful only in changing content, rather than fundamental pattern on their Rorschach psychograms (110). Tomkins, however, repeating the TAT daily with a group of subjects found that 20 sessions were required to bring out all the significant *themas* for one person (172). This finding is enough to indicate that high reliability can hardly be expected upon just one repetition of a test.

The most useful approach to the reliability of qualitative material is by means of comparisons between judges and interpreters. If a group of judges, using predetermined criteria for judgment, agree among themselves in the scoring of a number of protocols, reliability for the yardstick chosen is usually assumed. Stouffer obtained a reliability coefficient of .96 among four judges' estimates of attitudes toward prohibition based on subjects' autobiographies (266). In the application of this method the fact should not be overlooked that the coefficient may measure communality of thinking among the judges, resulting in spurious correspondence of ratings, quite independent of the particular materials to which the judgments are applied.

*Validity.* The most frequently proposed methods for establishing validity for projective techniques are (1) correspondence with other criteria, (2) internal consistency, and (3) predictive success. These are orthodox procedures which have long been advocated, but in their transfer to projective problems certain modifications are necessary due to the nature of the data which must be compared.

Clinical data have often been chosen as the reference point in validation studies. Many experiments of this sort have been conducted, only a few of which could be included in the bibliography of this paper. (Clinical validation of the Rorschach test is described in references 107, 116, 120, 126, 129, 133, and 137; references 145, 152, 153, 154, and 163 report similar experiments with the TAT.) In handling the material, the correlation technique has seldom been applicable, since it is impossible to reduce a complicated personality interpretation, or a life history, to a single variable for plotting on a chart. As a measure of congruence between data, Vernon's matching technique (271) has been used by Kerr (193), Murray (14), Hertz (115), and many others, with good results. For example, in the Harvard experiment already described, three judges were asked to match ten sets of phantasies with ten biographies. One judge matched five, and two matched all ten correctly (14, p. 390). Likewise, of 20 mosaic patterns done by ten normal and ten neurotic children in Kerr's experiment, 15 correct identifications

were made by the investigators. Contingency coefficients for various matchings in this study ranged from .86 to .96 (193).

There are two disadvantages in the matching technique. The first, which is shared by all validation procedures which depend upon correspondence with a criterion, is the difficulty of selecting an adequate standard. The usual procedure is to match a projective protocol with case history material, or other test results, but as Harrison (153) and Macfarlane (8) note, the case history itself may be highly inaccurate. The second objection is the dependence of the matching results upon the skill of judges. The judge who matched five would appear to possess less insight than the judge who matched ten in the experiment mentioned above. A poor judge can lower the matching coefficient while *true* congruence in the materials themselves remains unaltered, but still unknown. Regardless of these drawbacks, if successful matching can be accomplished, some degree of correspondence can safely be assumed and further tested by more refined measures.

The criterion of internal consistency, as originally applied to such personality inventories as the Thurstone and others which followed, has been used to test the agreement of separate items in a battery of questions with the battery as a whole. This method has been criticized on the grounds that validity thus established seldom holds beyond the standardization group. The unity demonstrated is between the given question and the original list of questions and does not constitute proof of coherence and "occurring-togetherness" of the traits themselves (254, p. 794).

In projective techniques, internal consistency applies to agreement found between results of assorted projective tests administered to the same subjects. But what constitutes *agreement*? The quotation from Frank cited above advocates this criterion; Allport lists it as one test for the validity of personal documents (209, p. 171); and Lerner and Murphy defend it as follows:

We aim for validity through comparison of a child with himself in different units of one and the same play situations in the nursery school and, insofar as possible at home. . . . The investigation of a few children, studied in this detailed fashion, can lead to as great or greater validity than large scale comparisons of so many children on a few ill-understood responses to *standard* stimuli (7. p. viii).

Although it seems reasonable to look for consistency within a person's performance in different situations, it must be recalled that true validity cannot be established through correspondence between measures which are not themselves validated. Even the conclusion that agreement offers evidence to support the theory that different tests measure the same thing (whatever it may be in the particular instance) calls for a warning sounded by Macfarlane. She points out that consistency may be a result of consistency in the experimenters' concepts, rather than between data themselves. (See discus-



sion of this point in connection with reliability in an earlier section of this paper.)

Probably the soundest tests of validity are tests of predictive capacity. In several studies of validity, indices based on prediction have proved to be higher than the reliability of the same ratings. Cartwright and French (223), after reading several years' diary entries, attempted to predict the diarist's answers to certain tests and questionnaires. Results showed more agreement between predicted and actual answers, than between the two judges' predictions. A similar phenomenon occurred when three experts made blind analyses of the Rorschach record of a single subject (115). Both Hertz (116) and Allport (209) explain the paradox as due to the fact that each analyst understood correctly different aspects of the personalities under inspection.

Prediction of actual behavior in concrete situations has not so far been attempted, if the designation *actual behavior* excludes responses to attitude tests (184, 199), prognosis in mental disease (124), or facts expected in the follow-up cases (132, 153), and if *concrete* implies a particular stimulus set-up. The possibility of suicide has been prognosticated from Rorschach records, the presence of hostile feelings and the likelihood of their overt expression has been noted, and numbers of similar motivational directions have been mapped and tested for individuals. It is not, however, possible to predict precise response sequences (Murray's *actones*, *verbones*, or *motones* [14]), since a diversity of acts may be equivalent for discharging the same drives in different individuals, and identical behavior may vary in motivation from person to person. Furthermore, a quantitative statement as to "how much" of a stimulus is needed to activate a tendency must remain relative rather than absolute.

*Standardization.* Progress toward the quantitative standardization of projective methods has been blocked not merely by the difficulties discussed above, but by a genuine distrust of the psychostatistical approach on the part of many. Allport illustrates the ludicrous results of what he terms "empiricism gone wild" by an example from one factor-analyzed inventory which assigns a score of plus six on *loyalty to the gang* for the response *green* to the word *grass* (210, p. 329). Frank points out that statistical tests of reliability and validity were originally devised to meet these problems in the absence of supplementary data (5, p. 400). Now that the latter are available, routine procedures should not be made a fetish to stand in the way of more effective, non-statistical techniques. Moreover, it has been claimed that *psychostatistical manipulations* and rigidly objective procedures are less applicable when carried over from the investigation of cognitive functions, such as intelligence, to the more affective aspects of total personality (154). *Paper and pencil* tests have come in for much unfavorable comment. Krugman

describes them as having had "high reliability but practically no validity when the criterion of validity was not an artificially constructed one but agreement with other clinical data" (128, p. 100). Harrison describes the same measures as possessing strict objectivity and high reliability, but as "lacking in the one requirement of a good test and that unfortunately rather an essential one, validity" (154, p. 49).

As an alternative, many investigators hold to the belief that, instead of being content with relatively invalid, objective tests, the more difficult and important task of validating subjective methods should be attempted. Harrison states this aim and adds:

In the end it may turn out to be an easier task to objectify a plastic, subjective method which is based on sound principles, than to validate an objective test standardized along lines of test construction which have time and again proved unsuccessful (152, p. 122).

Munroe sounds the same note in discussing the possible standardization of her *inspection technique* for the Rorschach. That this would be possible is quite evident from her results, but Munroe sees the real issue as whether or not it would be desirable.

Such standardization by its very nature ignores the individual. . . . None of our present personality scales, including the Rorschach, if it were standardized in this manner, rest upon anything more than an empirical approximation. All our theories of personality are at variance with the notion that the summation of a series of items determined by discrete frequency tables could ever be expected to give an accurate dynamic picture of an individual (130, p. 233).

*Criticisms and cautions.* The opinion quoted so far in this survey has emanated almost entirely from within the frame of reference of projective techniques, rather than from the critical vantage point of a detached observer. What of the other side? What have the much criticized, even belittled adherents to traditional methodology in the personality field to say on the subject? Unfortunately, although there are many who disapprove the trend, projective methods have for the most part been "damned with faint praise" or simply ignored. Critics of the Rorschach, for example, are known to be numerous, but experimental refutations of claims made are practically nonexistent. There is, in the literature, not one comprehensive critical review of projective methods by anyone outside the group of interested researchers. Does this mean that the advocates of newer approaches are attacking insensible straw men, or are they alive but inarticulate? Even the ESP research was for a time honored by highly vocal opposition.

Only two frankly critical articles can be cited. One was read by Balken at the 1941 meeting of the American Psychological Association (1). This paper



directed its attack chiefly at the lack of precision in the terminology used, and against what the author regarded as unjustified claims and interpretations, but which she did not specifically illustrate. In further discussion of her paper at a round-table meeting, Balken expressed objections which appeared to be grounded in the belief that psychoanalysis has been insufficiently credited for underlying principles, and that its theories are being irresponsibly extended without sufficient knowledge of its basic implications.

Macfarlane (8) discusses the problems of standardization, reliability and validity, and adds two important cautions. The first relates to sampling. Clinicians, she warns, are "conspicuously subject to faulty over-generalization" and should be required to tabulate their own sample experience as a check against extending theories beyond the groups familiar to the investigator. She speaks of the *mushroom growth* of the methods, objects to their use by persons of sundry qualification or lack of it, and pleads that the inexperienced should be barred from this area of research, lest promising leads be bungled. Furthermore, she emphasizes the important and often neglected fact that all validation must rest upon the concepts and hypotheses underlying the research. Since there are as yet no universally accepted terms which do justice to the *richness and diversity* of personality, she suggests that an *articulated conceptualization* is basic.

It should become a part of the scientific mores in this pioneering, unstructured stage, that the first step in projective research should be an explicit statement of concepts used, and an orientation with respect to theoretical biases. Further, such a statement should appear on page one of any article instead of leaving it to the inference of the reader (8, p. 406).

Although projective methods can be and frequently are used for nomothetic as well as idiographic purposes, they are peculiarly well suited to the former; hence the controversy which rages about the single case study is an important related issue. John E. Anderson objects that the single case in physics in which the weight of factors is known in advance and controlled in order to design a crucial experiment is not comparable to the single case in psychology in which there is no knowledge of the weight of factors and little or no possibility of control. This criticism, although pertinent in its emphasis upon the absence of crucial experimentation in the psychology of personality, appears to arise from a difference in fundamental premise. If we regard personality as an aggregate of parts, their action and interaction are obviously too complex for control. No matter how standardized the external situation, we cannot control the momentary inner state of the individual due to influences beyond our jurisdiction. If, on the other hand, we join the organismic personologists in regarding personality as a unified

whole, then the person and not the part becomes the unit for study. This unit may be thought of as a construct, no more and no less directly observable than an atom. From this viewpoint it is feasible to control the immediate external conditions of experiment, and to study the composition of our *substance* (that elusive *inner state* or those *processes within* which baffle other procedures) through its effect upon something else, as in the physical experiments Frank describes. Exponents of the totalistic approach claim more logic and greater rigor in such method, than in attempts to set up elaborate control of conditions surrounding some arbitrary partial aspect of personality (such as a trait or an attitude) which is rendered distorted and meaningless by removal from the intrapersonal context.

Certain further disadvantages of projective methods have been noted by interested workers within the field. Some of the techniques, such as the *Tautaphone*, are considered better suited to research than to clinical diagnosis because they are awkward, time consuming, and add little information which could not be more efficiently obtained by a psychiatric interview (174). Harrison holds that, except for the Rorschach and the TAT, which have been *partially validated*, most of the methods are *terra incognita* with validity and true value still to be demonstrated. "Others appear best suited for qualitative and interpretive work, and by their nature offer little opportunity for compromise with objective procedures" (154, p. 52).

### *Evaluation*

A review of the literature on projective methods shows first that they have a long and respectable lineage, with ties of kinship extending beyond the boundaries of psychology itself. Whether the offspring will grow into ne'er-do-well dreamers, fulfilling the prophecies of conservative relatives, or will develop their originality in a maturity of scientific respectability, is a matter for the future to decide.

The theoretical issues will probably never be settled in an *either-or* fashion, since they represent philosophies of science which, in one shape or another, are as old as science itself. Psychology does well to be jealous of its painstaking empirical techniques, in spite of scoffers who hold that "the more exact methods generally yield the least information" (14, p. 547). On the other hand, if complex molar behavior and its hidden springs of action are not to be ignored, an open-minded attitude toward theory revision and toward innovations in method seems necessary. The admission of *explanatory theory*, combined with attempts to verify its constructs by experimentation, has been offered as one solution.



Bavink, the philosopher, has distinguished two types of theory: *elaborative* theory which contains practically no hypothetical elements since the fundamental assumptions are themselves data of experience; and *explanatory* theory in which "the unification of facts is only reached on the basis of a speculative assumption which is described as the hypothesis on which the theory is based," as in the atomic theory (215, p. 168).

Certain immediate issues seem more pseudo than real. Does standardization, for example, actually involve forsaking such important aspects as context and interrelationships among variables? If so, perhaps it is reasonable to ask that it be postponed or even abandoned. Certainly normative data, interindividual comparisons on single traits, and the search for uniformities should not be pursued exclusively at the expense of intensive individual studies of both the horizontal and vertical type. Interest in the methodology of personal documents and intensive projects such as the Harvard and Sarah Lawrence studies help to correct overemphasis. But does the value of the new approach rule out the need for normative research as well? It would seem that the latter could at least provide orienting data which, properly used and interpreted, could furnish the clinician and student of personality with a stronger framework for his studies in interrelationship.

A fact which is often overlooked both by those who scorn statistics and those who reify them is that numbers do not bestow precision but are, rather, a convenient way to express it when it exists. Quantitative method might profitably be applied more extensively to the properties of the projective tests themselves. For example, as Harrison suggests (154), it would be valuable to have frequency tables similar to those for the Kent Rosanoff word-association test, showing the relative frequency of certain common phantasies produced by each of the TAT pictures. If we think of these numerical results as adding to the precision of the instrument itself, instead of reading into the figures oversimplified generalizations about people, no atomistic conception of personality is implied. The clinician would not need to alter either his theory or his interpretation of certain Rorschach syndromes if he also knew more about the frequency of the component determinants, both singly and in constellations. Such knowledge would, on the contrary, serve both as an added support and as a check on his conclusions.

Beck (103), Hertz (117) and others have developed tables showing the most commonly perceived forms on the Rorschach test. Zubin (139-141) has suggested a radical revision of Rorschach scoring to provide for more exact quantification. Common or *normal* details selected; forms frequently seen in the blots; as well as response determinants such as color, shading and movement have been intensively studied (106, 117, 118, 121, 124). Nothing comparable has been done for some of the other aspects of scoring,

and rigid tests of various ratios used in interpretation have not been carried out. Other projective techniques are still more difficult to quantify and little or nothing has been done to develop workable scoring schemes.

Questions raised in regard to sampling are another issue which ought not to be insoluble. If projective methods are used to provide comparative data, Macfarlane's caution (see above) should be observed. For situations of this kind, the accepted rules of random sampling are sufficient, but it may be that we need supplementary principles to determine the adequacy of sampling in individual studies. An individual, as F. H. Allport has suggested (208), may also be regarded as a *population* of *events* and characteristics. What constitutes an adequate sample from an individual life? The validity of individual diagnosis and prediction might be increased by controls relating to the number of tests to be given, their distribution in time or under varying emotional conditions, the number of different observers needed, and so on. Some of these problems would be difficult to investigate, but with ingenuity should not be impossible to attack. For example, Tomkins' daily repetition of the TAT to determine the new thema limit (172) is one step in the direction of amassing data which might be used as a basis for individual sampling theory.

Recent advances in statistics have been little exploited in the quantitative treatment of results of projective experiments capable of expression in numbers, whether derived from repeated observations on single individuals under systematically varied conditions or from multivariable experiments involving exhaustive treatment of a small number of cases. Many authors are content to report their findings in per cents without statements of significance. Small-sample theory, including the techniques of analysis of variance and covariance have not been used to any extent in projective experimentation, although Brenman and Reichardt (109) have recently made use of Fisher's *t*-test to show the significance of differences in hypnotizability predicted from Rorschach records.

The advantage of these statistics, originally developed by Fisher in agricultural research and adapted to educational problems by Snedecor (264) and Lindquist (250), lies not in the unfortunate assumption that they offer a means for obtaining *significant* figures from scanty data or carelessly designed experiments. Instead, they avoid generalization beyond the data at hand by supplying a numerical means of expressing relationships found in a sample and of comparing these with results expected by chance in *samples of the same size* (250, p. 54). Moreover, the poorly designed experiment tends to defeat itself, since inadequate control of important variables results in large errors (*Within groups* variances) which reduce significance. Three advantages of the variance methods which seem especially promising as applied to projective experiments are as follows: First, in the exploratory study in which a number of factors must be controlled, these can be



handled simultaneously. Second, by the selection of cases to fit prescribed conditions, variables may be controlled without the laborious necessity of holding them constant. Finally, a measure of the amount and significance of variability due to the interaction of one or more variables, or to the effect of uncontrolled variables, is obtainable (232, 233).

For clinical psychology, which may be regarded as the applied branch of the psychology of personality, projective methods furnish one of the most promising hopes for a science of diagnosis and treatment. An experimental study by Davis (227) on the relative weight given to data from tests and from case histories in arriving at clinical judgments found that clinicians are unwilling and unable to rely solely on objective criteria. Many believe that the final synthesis must forever rest upon the clinician's skill, experience, and "intuition," but the fact that these assets must be possessed and used does not relieve us of the responsibility of attempting at every turn to expand such subjective equipment by the development of scientific devices to supplement and check upon our conjectures.

Murray has remarked that psychology has the choice of two alternatives: to study important problems with as yet inadequate instruments, or to study with adequate instruments unimportant problems (15). Whatever their present status, the interest in projective procedures is evidence of an attempt to improve upon inadequate techniques, and evinces a constructive and hopeful preoccupation with method itself. There is a growing recognition that we must not be limited by narrow conceptions of what constitutes *the* scientific method, but instead must be on the watch for new approaches which conform to the broad aims of science, rather than to its dogmas (211). Historically, science arose in response to the need for knowing and dealing with the natural world, or in other words:

Science aims to give man an understanding, a power of prediction, and a power of control, beyond that which he can achieve through his own unaided common sense (209, p. 148).

If projective methods can be refined and safeguarded in order to serve that end, they deserve interested attention and exhausted research. There appears to be considerable evidence that they may be well worth the expenditure of time and effort involved in thorough exploration.

## BIBLIOGRAPHY

Since the term *projective methods* includes a variety of techniques, a number of which represent special interest areas, the bibliography has been organized accordingly. Section I lists articles devoted chiefly to theory, or to discussions and experi-

ments which include an assortment of representative techniques. Section II covers all of the special methods, subdivided alphabetically under the appropriate headings. In Section III will be found books and articles of historical importance, and other works mentioned in the text as having theoretical or incidental bearing on the problems involved. The sections on *Projective Methodology*, *Thematic Apperception*, and *Miscellaneous Projective Techniques* are believed to be nearly exhaustive at the time of writing. The Rorschach section is selective since a complete list would run to hundreds of titles. The section on *Play techniques* omits a large number of articles concerned with their use in therapy.

## SECTION I

### *Projective Methodology*

1. BALKEN, EVA RUTH. Projective techniques for the study of personality. A critique. *Psychol. Bull.*, 1941, 38, 596. (Abstract.)
2. BROWN, J. F. & RAPAPORT, D. The role of the psychologist in the psychiatric clinic. *Bull. Menninger Clin.*, 1941, 5, 75-84.
3. BOOTH, GEORGE C. Objective techniques in personality testing. *Arch. Neurol. Psychiat.*, Chicago, 1939, 42, 514-530.
4. FEIGENBAUM, D. On projection. *Psychoanal. Quart.*, 1936, 5, 303-319.
5. FRANK, L. K. Projective methods for the study of personality. *J. Psychol.*, 1939, 8, 389-413. (Also in *Trans. N. Y. Acad. Sci.*, 1939, 1, 129-132.)
6. HOROWICZ, RUTH, & MURPHY, LOIS. Projective methods in the psychological study of children. *J. Exp. Educ.*, 1938, 7, 133-140.
7. LERNER, E., MURPHY, LOIS, STONE, L. J., BEYER, EVELYN, & BROWN, ELINOR. Studying child personality. *Monogr. Soc. Res. Child Develpm.*, 1941, 6, No. 4.
8. MACFARLANE, JEAN W. Problems of validation inherent in projective methods. *Amer. J. Orthopsychiat.*, 1942, 12, 405-411.
9. MASLOW, A. H. & MITTELMANN, B. *Principles of abnormal psychology*. Appendix I. Projective methods of examination. Pp. 611-622. New York: Harper, 1941.
10. MOELLENHOFF, F. A projection returns and materializes. *Amer. Imago*, 1942, 3, 3-13.
11. MURPHY, LOIS B. Interiorization of family experiences by normal preschool children as revealed by some projective methods. *Psychol. League J.*, 1940, 4, 3-5.
12. MURPHY, LOIS B. Patterns of spontaneity and constraint in the use of projective materials by preschool children. *Trans. N.Y. Acad. Sci.*, 1942, 4, 124-138.
13. MURRAY, H. A. Effect of fear on estimates of maliciousness of other personalities. In Tomkins, Silvan S. *Contemporary Psychopathology*. Cambridge: Harvard Univ. Press, 1943. Pp. 545-561.
14. MURRAY, H. A. *Explorations in personality*. New York: Oxford Univ. Press, 1938.
15. MURRAY, H. A. An investigation of fantasies. In Abstract of C. L. Hull's informal seminar at Yale University, 1936. (Unpublished.)
16. RAPAPORT, D. Principles underlying projective techniques. *Character & Pers.*, 1942, 10, 213-219.



17. RICHENBERG, W. & CHIDESTER, LEONA. Lack of imagination as a factor in delinquent behavior. *Bull. Menninger Clin.*, 1937, 1, 226-231.
18. ROSENZWEIG, S. Fantasy in personality and its study by test procedures. *J. Abnorm. Soc. Psychol.*, 1942, 37, 40-51.
19. SEARS, R. R. Experimental studies of projection: I. Attribution of traits. In Tomkins, Silvan S. *Contemporary Psychopathology*, Cambridge: Harvard Univ. Press, 1943. Pp. 561-571.
20. STRANG, RUTH. Technical instruments of mental hygiene diagnosis and therapy. *Rev. Educ. Res.*, 1940, 10, 450-459.
21. SYMONDS, P. M. & SAMUEL, E. A. Projective methods in the study of personality. *Rev. Educ. Res.*, 1941, 11, 80-93.
22. UPDEGRAFF, RUTH. Recent approaches to the study of the preschool child. I. Indirect and projective methods. *J. Consult. Psychol.*, 1938, 2, 159-161.
23. WHITE, R. W. The interpretation of imaginative productions. In Hunt, J. McV. *Personality and the Behavior Disorders. A handbook based on experimental research.* (2 vol.) New York: Ronald Press, 1944. Pp. 214-254.
24. WOLFF, W. Projective methods for personality analysis of expressive behavior in preschool children. *Character & Pers.*, 1942, 10, 309-330.

## SECTION II

*Art Techniques*

25. ABEL, THEODORA M. Free designs of limited scope as a personality index. *Character & Pers.*, 1938, 7, 50-62.
26. ANASTASI, A. & FOLEY, J. P. A survey of the literature on artistic behavior in the abnormal: III. Spontaneous productions. *Psychol. Monogr.*, 1942, 52, No. 237.
27. BARNHART, E. N. Stages in construction of children's drawing as revealed through a recording device. *Psychol. Bull.*, 1940, 37, 581. (Abstract.)
28. BENDER, LAURETTA. Art and therapy in the mental disturbances of children. *J. Nerv. Ment. Dis.*, 1937, 86, 249-263.
29. BENDER, LAURETTA. Gestalt principles in the side-walk drawings and games of children. *J. Genet. Psychol.*, 1932, 41, 192-210.
30. BENDER, LAURETTA. The Goodenough test in chronic encephalitis in children. *J. Nerv. Ment. Dis.*, 1940, 91, 277-286.
31. BENDER, LAURETTA & WOLTMANN, A. G. The use of plastic material as a psychiatric approach to emotional problems of children. *Amer. J. Orthopsychiat.*, 1937, 7, 283-300.
32. FLEMING, J. Observations on the use of finger painting in the treatment of adult patients with personality disorders. *Character & Pers.*, 1940, 8, 301-310.
33. HARMS, E. Child art as an aid in the diagnosis of juvenile neuroses. *Amer. J. Orthopsychiat.*, 1941, 11, 191-209.
34. LEVY, J. Use of art techniques in treatment of children's behavior problems. *Proc. Amer. Ass. Stud. Ment. Def.*, 1934, 58, 258-260.

35. LEWIS, N. D. C. Graphic art productions in schizophrenia. *Proc. Ass. Res. Nerv. ment. Dis.*, 1928, 5, 344-368.
36. MCINTOSH, J. R. An inquiry into the use of children's drawings as a means of psychoanalysis. *Brit. J. Educ. Psychol.*, 1939, 9, 102-103. (Abstract.)
37. MCINTOSH, J. R. & PICKFORD, R. W. Some clinical and artistic aspects of a child's drawings. *Brit. J. Med. Psychol.*, 1943, 19, 342-362.
38. MOSSE, E. P. Painting analysis in the treatment of neuroses. *Psychoanal. Rev.*, 1940, 27, 68-82.
39. NAUMBERG, M. Children's art expressions and war. *Nerv. Child*, 1943, 2, 360-373.
40. REITMAN, F. Facial expression in schizophrenic drawing. *J. Ment. Sci.*, 1939, 85, 264-272.
41. SCHMIDLE-WAEHNER, T. Formal criteria for the analysis of children's drawing. *Amer. J. Orthopsychiat.*, 1942, 12, 95-104.
42. SCHUBE, K. & COWELL, G. Art of psychotic persons. *Arch. Neurol. & Psychiat.*, Chicago, 1939, 41, 709-720.
43. SHAW, RUTH F. *Finger painting*. Boston: Little Brown, 1934.
44. SHAW, RUTH F. & LYLE, JEANNETTE. Encouraging fantasy expression in children. *Bull. Menninger Clin.*, 1937 1, 78-86.
45. SPOERL, D. T. Personality and drawing in retarded children. *Character & Pers.*, 1940, 8, 227-239.
46. SPRINGER, N. N. A study of the drawings of maladjusted and adjusted children. *J. Genet. Psychol.*, 1941, 58, 131-138.
47. WILLIAMS, J. N. Interpretation of drawings made by maladjusted children. *Virg. Med. Monogr.*, 1940, 67, 533-538.

### *Drama and Puppets*

48. BENDER, LAURETTA, & WOLTMANN, A. G. The use of puppet shows as a psychotherapeutic method for behavior problems in children. *Amer. J. Orthopsychiat.*, 1936, 6, 341-354.
49. BORDIN, RUTH. The use of psychodrama in an institute for delinquent girls. *Sociometry*, 1940, 3, 80-90.
50. CURRAN, FRANK J. The drama as a therapeutic measure in adolescence. *Amer. J. Orthopsychiat.*, 1939, 9, 215-232.
51. FRANZ, J. G. The place of the psychodrama in research. *Sociometry*, 1940, 3, 49-61.
52. JENKINS, R. L. & BECKH, E. Finger puppets and mask making as a media for work with children. *Amer. J. Orthopsychiat.*, 1942, 12, 294-301.
53. LYLE, J. & HOLLY, S. B. The therapeutic value of puppets. *Bull. Menninger Clin.*, 1941, 5, 223-226.
54. MORENO, J. L. Mental catharsis and the psychodrama. *Sociometry*, 1940, 3, 209-244.
55. MORENO, J. L. Psychodramatic shock therapy. A sociometric approach to the problem of mental disorders. *Sociometry*, 1939, 2, 1-30.



56. WOLTMANN, A. G. The use of puppets in understanding children. *Ment. Hyg., N.Y.*, 1940, 24, 445-458.

### *Play Techniques*

57. ACKERMAN, N. W. Constructive and destructive tendencies in children. *Amer. J. Orthopsychiat.*, 1937, 7, 301-319.
58. ACKERMAN, N. W. Constructive and destructive tendencies in children. An experimental study. *Amer. J. Orthopsychiat.*, 1938, 8, 265-285.
59. ALLEN, F. H. *Psychotherapy with children*. New York: W. W. Norton, 1942.
60. ALLEN, F. H. Therapeutic work with children. *Amer. J. Orthopsychiat.*, 1934, 4, 193-202.
61. BARUCH, DOROTHY W. Aggression during doll play in a preschool. *Amer. J. Orthopsychiat.*, 1941, 11, 252-260.
62. BARUCH, DOROTHY W. Play techniques in pre-school as an aid in guidance. *Psychol. Bull.*, 1939, 36, 570. Abstract.
63. BENDER, LAURETTA & SCHILDER, P. Aggressiveness in children. *Genet. Psychol. Monogr.*, 1936, 18, 410-525.
64. BENDER, LAURETTA & SCHILDER, P. Form as a principle in the play of children. *J. Genet. Psychol.*, 1936, 49, 254-261.
65. CAMERON, W. M. The treatment of children in psychiatric clinics with particular reference to the use of play techniques. *Bull. Menninger Clin.*, 1940, 4, 172-180.
66. CONN, J. H. The child reveals himself through play. *Ment. Hyg., N.Y.*, 1939, 23, 49-69.
67. CONN, J. H. The play interview. A method of studying children's attitudes. *Amer. J. Dis. Child.*, 1939, 58, 1199-1214.
68. CONN, J. H. A psychiatric study of car sickness in children. *Amer. J. Orthopsychiat.*, 1938, 8, 130-141.
69. DESPERT, J. LOUISE. A method for the study of personality reactions in pre-school age children by means of analysis of their play. *J. Psychol.*, 1940, 9, 17-29.
70. DESPERT, J. LOUISE. Technical approaches used in the study of emotional problems in children. Part IV. Collective phantasy. *Psychiat. Quart.*, 1937, 11, 491-506.
71. DESPERT, J. LOUISE. Technical approaches used in the study of emotional problems in children. Part V. The playroom. *Psychiat. Quart.*, 1937, 11, 677-693.
72. ERIKSON, ERIK HOMBURGER. Configurations in play. Clinical notes. *Psychoanal. Quart.*, 1937, 6, 139-214.
73. ERIKSON, ERIK HOMBURGER. Further explorations in play construction. Three spatial variables in relation to sex and anxiety. *Psychol. Bull.*, 1938, 41, 748. (Abstract.)
74. ERIKSON, ERIK HOMBURGER. Studies in the interpretation of play. Clinical observations of play disruption in young children. *Genet. Psychol. Monogr.*, 1940, 22, 557-671.

75. FREUD, ANNA. Introduction to the technique of child analysis. (Authorized trans. supervised by L. P. Clark.) *Nerv. & Ment. Dis. Monogr.*, 1928, No. 48.
76. FRIES, MARGARET E. Play technique in the analysis of young children. *Psychoanal. Rev.*, 1937, 24, 233-245.
77. FRIES, MARGARET E. The value of play for a child development study. *Understand. Child*, 1938, 7, 15-18.
78. GITELSON, M. Clinical experience with play therapy. *Amer. J. Orthopsychiat.*, 1938, 8, 466-478.
79. GITELSON, M., ROSS, H., HOMBURGER, E., ALLEN, F., BLANCHARD, PHYLLIS, LIPPMAN, H. S., GERARD, M. & LOWRY, L. Section on play therapy. *Amer. J. Orthopsychiat.*, 1938, 8, 499-524.
80. HOLMER, P. The use of the play situation as an aid to diagnosis. A case report. *Amer. J. Orthopsychiat.*, 1937, 7, 523-531.
81. KANNER, LEO. Play investigations and play treatment of children's behavior disorders. *J. Pediat.*, 1940, 17, 533-545.
82. KLEIN, MELANIE. The psychoanalysis of children. London: Hogarth Press, 1932.
83. LEVY, D. Studies in sibling rivalry. *Res. Mongr. Amer. Orthopsychiat. Ass.*, 1937, No. 2.
84. LEVY, D. Use of play technique as experimental procedure. *Amer. J. Orthopsychiat.*, 1933, 3, 266-277.
85. LISS, E. Play techniques in child analysis. *Amer. J. Orthopsychiat.*, 1936, 6, 17-22.
86. LOWENFELD, MARGARET. The theory and use of play in the psychotherapy of childhood. *J. Ment. Sci.*, 1938, 4, 1057-1058.
87. MAYER, A. M. & MAYER, E. B. Dynamic concept test. A modified play technique for adults. *Psychiat. Quart.*, 1941, 15, 621-634.
88. NEWELL, H. W. Play therapy in child psychiatry. *Amer. J. Orthopsychiat.*, 1941, 11, 245-252.
89. RICHARDS, S. S. & WOLFF, E. The organization and function of play activities in the set-up of a pediatrics department. *Ment. Hyg., N. Y.*, 1940, 24, 229-235.
90. ROGERSON, C. H. *Play therapy in childhood*. New York: Oxford Univ. Press, 1939.
91. ROSENZWEIG, S. & SHAKOW, D. Play techniques in schizophrenia and other psychoses. I. Rationale. *Amer. J. Orthopsychiat.*, 1937, 7, 32-35.
92. ROSENZWEIG, S. & SHAKOW, D. Play techniques in schizophrenia and other psychoses. II. Schizophrenic constructions. *Amer. J. Orthopsychiat.*, 1937, 7, 36-47.
93. SARGENT, HELEN D. Spontaneous doll play of a nine-year-old. *J. Consult. Psychol.*, 1943, 7, 216-222.
94. SIMPSON, G. Diagnostic play interviews. *Understand. Child*, 1938, 7, 6-10.
95. SOLOMON, J. C. Active play therapy. *Amer. J. Orthopsychiat.*, 1938, 8, 479-498.
96. SOLOMON, J. C. Active play therapy. Further experiences. *Amer. J. Orthopsychiat.*, 1940, 10, 763-781.
97. SYMONDS, P. M. Play technique as a test of readiness. *Understand. Child*, 1940, 9, 8-14.



98. TALLMAN, F. & GOLDENSOHN, L. N. Play techniques. *Amer. J. Orthopsychiat.*, 1941, 11, 551-561.
99. WALDER, R. The psychoanalytic theory of play. *Psychoanal. Quart.*, 1933, 2, 208-224.
100. WEISS-FRANKL, A. B. Diagnostic and remedial play. *Understand. Child*, 1938, 7, 3-5.

### *Rorschach Method*

101. BECK, S. J. Configurational tendencies in Rorschach responses. *Amer. J. Psychol.*, 1933, 45, 433-443.
102. BECK, S. J. Error, symbol and method in the Rorschach test. *J. Abnorm. Soc. Psychol.*, 1942, 37, 83-103.
103. BECK, S. J. Introduction to the Rorschach method. A manual of personality study. *Res. Monogr. Amer. Orthopsychiat. Ass.*, 1937, No. 1.
104. BECK, S. J. Personality structure in schizophrenia. A Rorschach investigation on 81 patients and 64 controls. *Nerv. Ment. Dis. Monogr.*, 1938, No. 63.
105. BECK, S. J. The Rorschach test in psychopathology. *J. Consult. Psychol.*, 1943, 7, 103-111.
106. BECK, S. J. *Rorschach's test*. New York: Grune & Stratton, 1944. Vol. I. Elementary principles.
107. BENJAMIN, J. D. & EBAUGH, F. G. The diagnostic validity of the Rorschach test. *Amer. J. Psychiat.*, 1938, 94, 1163-1178.
108. BOCHNER, RUTH & HALPERN, FLORENCE. *Clinical application of the Rorschach test*. New York: Grune & Stratton, 1942.
109. BRENNAN, MARGARET & REICHARD, SUZANNE. Use of the Rorschach test in the prediction of hypnotizability. *Bull. Menninger Clin.*, 1943, 7, 183-188.
110. FOSBERG, I. A. An experimental study of the reliability of the Rorschach psychodiagnostic technique. *Rorschach Res. Exch.*, 1941, 5, 72-84.
111. FRANK, L. K. Foreword to issue on the Rorschach method. *J. Consult. Psychol.*, 1943, 7, 63-66.
112. HARROWER-ERICKSON, MOLLIE R. The contribution of the Rorschach method to war-time psychological problems. *J. Ment. Sci.*, 1940, 86, 1-12.
113. HARROWER-ERICKSON, MOLLIE R. Large scale experimentation with the Rorschach method. *J. Consult. Psychol.*, 1943, 7, 120-127.
114. HARROWER-ERICKSON, MOLLIE R. Modification of the Rorschach method for use as a group test. *Rorschach Res. Exch.*, 1941, 5, 130-144.
115. HERTZ, MARGUERITE R. Comparison of three blind Rorschach analyses. *Amer. J. Orthopsychiat.*, 1939, 9, 295-315.
116. HERTZ, MARGUERITE R. Rorschach twenty years after. *Psychol. Bull.*, 1942, 39, 529-572.
117. HERTZ, MARGUERITE R. Scoring the Rorschach test with specific reference to the normal detail category. *Amer. J. Orthopsychiat.*, 1938, 8, 100-121.

118. HERTZ, MARGUERITE, R. The shading response in the Rorschach inkblot test: a review of its scoring and interpretation. *J. Gen. Psychol.*, 1940, 23, 123-167.
119. HERTZ, MARGUERITE R. The Rorschach method: science or mystery. *J. Consult. Psychol.*, 1943, 7, 67-80.
120. HERTZ, MARGUERITE R. Validity of the Rorschach method. *Amer. J. Orthopsychiat.*, 1941, 11, 512-520.
121. HERTZMANN, M. Recent research on the group Rorschach test. *Rorschach Res. Exch.*, 1943, 7, 1-6.
122. HERTZMANN, M. & MARGULIES, HELEN. Developmental changes in Rorschach test responses. *J. Genet. Psychol.*, 1943, 62, 189-216.
123. KELLEY, D. M. The present state of the Rorschach method as a psychological adjunct. *Rorschach Res. Exch.*, 1940, 4, 30-36.
124. KISKER, G. W. A projective approach to personality patterns during insulin shock and metrazol convulsive therapy. *J. Abnorm. Soc. Psychol.*, 1942, 37, 120-124.
125. KLOFFER, B. & KELLY, D. *The Rorschach technique*. Yonkers on Hudson: World Book, 1942.
126. KRUGMAN, J. E. A clinical validation of the Rorschach with problem children. *Rorschach Res. Exch.*, 1942, 6, 61-70.
127. KRUGMAN, M. The Rorschach in child guidance. *J. Consult. Psychol.*, 1943, 7, 80-88.
128. KRUGMAN, M. Out of the inkwell: the Rorschach method. *Character & Pers.*, 1940, 9, 91-110.
129. MIALE, F. R., CLAPP, H. & KAPLAN, A. H. Clinical validation of a Rorschach interpretation. *Rorschach Res. Exch.*, 1938, 2, 153-163.
130. MUNROE, RUTH. An experiment in large scale testing by a modification of the Rorschach method. *J. Psychol.*, 1942, 13, 229-263.
131. MUNROE, RUTH. The inspection technique. A modification of the Rorschach method of personality diagnosis for large scale application. *Rorschach Res. Exch.*, 1941, 5, 166-190.
132. MUNROE, RUTH. Use of the Rorschach in college counseling. *J. Consult. Psychol.*, 1943, 7, 89-97.
133. PIATROWSKI, Z. Blind analysis of a case of compulsion neurosis. *Rorschach Res. Exch.*, 1937, 2, 89-111.
134. PIATROWSKI, Z. Use of the Rorschach in vocational selection. *J. Consult. Psychol.*, 1943, 7, 97-102.
135. RORSCHACH, H. *Psychodiagnostics: a diagnostic test based on perception*. (Trans. by P. Lemkau & B. Kronenburg.) Bern: Hans Huber, 1942. New York: Grune & Stratton.
136. RORSCHACH, H. & OBERHOLZER, E. The application of the interpretation of form to psychoanalysis. *J. Nerv. Ment. Dis.*, 1924, 60, 225-248; 359-379.
137. VARVEL, W. A. Suggestions toward the experimental validation of the Rorschach test. *Bull. Menninger Clin.*, 1937, 1, 220-226.



138. VERNON, P. E. The Rorschach inkblot test. *Brit. J. Med. Psychol.*, 1933, 13, 89-118; 179-200; 271-291.
139. ZUBIN, J. A. Psychometric approach to the evaluation of the Rorschach test. *Psychiatry*, 1941, 4, 547-566.
140. ZUBIN, J. A. Quantitative approach to measuring regularity of succession in the Rorschach experiment. *Character & Pers.*, 1941, 10, 67-78.
141. ZUBIN, J. A., CHUTE, E. & VERNIAR, E. Psychometric scales for scoring Rorschach test responses. *Character & Pers.*, 1943, 11, 277-301.
142. YOUNG, R. A. & HIGGENBOTHAM, S. A. Behavior checks on the Rorschach method. *Amer. J. Orthopsychiat.*, 1942, 12, 87-95.

### *Thematic Apperception*

143. AMEN, E. W. Individual differences in apperceptive reaction: a study of response of pre-school children to pictures. *Genet. Psychol. Monogr.*, 1941, 23, 319-385.
144. BALKEN, EVA RUTH. A delineation of schizophrenic language and thought in a test of imagination. *J. Psychol.*, 1943, 16, 239-272.
145. BALKEN, EVA RUTH & MASSERMAN, J. H. The language of phantasy. III. The language of phantasies of patients with conversion hysteria, anxiety state and obsessive compulsive neuroses. In Tomkins, Silvan S. *Contemporary Psychopathology*. Cambridge: Harvard Univ. Press, 1943, Pp. 244-253.
146. BALKEN, EVA RUTH & VANDERVEER, A. H. The clinical application of the Thematic apperception test to neurotic children. *Psychol. Bull.*, 1940, 37, 517. (Abstract.)
147. BALKEN, EVA RUTH & VANDERVEER, A. H. The clinical application of a test of imagination to neurotic children. *Amer. J. Orthopsychiat.*, 1942, 12, 68-81.
148. BELLAK, L. An experimental investigation of projection. *Psychol. Bull.*, 1942, 39, 489. (Abstract.)
149. BENNETT, GEORGIA. Structural factors related to the substitute value of activities in normal and schizophrenic persons. I. A technique for the investigation of central areas of personality. *Character & Pers.*, 1941, 10, 42-50.
150. BENNETT, GEORGIA. Some factors related to substitute value at the level of fantasy. *Psychol. Bull.*, 1942, 39, 488. (Abstract.)
151. CHRISTENSON, J. A. JR. Clinical application of the Thematic apperception test. *J. Abnorm. Soc. Psychol.*, 1943, 38, 104-107.
152. HARRISON, R. Studies in the use and validity of the Thematic apperception test with mentally disordered patients. II. A quantitative study. *Character & Pers.*, 1940, 9, 122-133.
153. HARRISON, R. Studies in the use and validity of the Thematic apperception test with mentally disordered patients. III. Validation by the method of "blind analysis." *Character & Pers.*, 1940, 9, 134-138.
154. HARRISON, R. The Thematic apperception and Rorschach methods of personality investigation in clinical practice. *J. Psychol.*, 1943, 15, 49-74.
155. MASSERMAN, J. H. & BALKEN, EVA RUTH. The clinical application of phantasy studies. *J. Psychol.*, 1938, 6, 81-88.

156. MASSERMAN, J. H. & BALKEN, EVA RUTH. The psychoanalytic and psychiatric significance of phantasy. *Psychoanal. Rev.*, 1939, 26, 243-279.
157. MORGAN, C. D. & MURRAY, H. A. A method for investigating fantasies: the Thematic apperception test. *Arch. Neurol. Psychiat.*, Chicago, 1935, 34, 289-306.
158. MURRAY, H. A. *Manual for the Thematic apperception test*. Cambridge: Harvard Univ. Press, 1943.
159. MURRAY, H. A. Techniques for a systematic investigation of fantasy. *J. Psychol.*, 1937, 3, 115-145.
160. RAPAPORT, D. The clinical application of the Thematic apperception test. *Bull. Menninger Clin.*, 1943, 7, 106-113.
161. RAPAPORT, D. The Thematic apperception test. Qualitative conclusions as to its interpretation. *Psychol. Bull.*, 1942, 39, 592. (Abstract.)
162. RODNICK, E. H. & KLETANOFF, S. G. Projective reactions to induced frustrations as a measure of social adjustment. *Psychol. Bull.*, 1942, 39, 389. (Abstract.)
163. ROTTER, J. R. Studies in the use and validity of the Thematic apperception test with mentally disordered patients. I. Method of analysis and clinical problems. *Character & Pers.*, 1940 9, 18-34.
164. SANFORD, R. N. *Procedure for scoring the Thematic apperception test*. Cambridge: Harvard Psychological Clinic, 1939. (Privately printed.)
165. SANFORD, R. N. Some quantitative results from the analysis of children's stories. *Psychol. Bull.*, 1941, 38, 749. (Abstract.)
166. SARASON, S. B. The use of the Thematic apperception test with mentally deficient children. I. A study of high grade girls. *Amer. J. Ment. Def.*, 1943, 47, 414-421.
167. SARASON, S. B. & ROSENZWEIG, S. An experimental study of the triadic hypothesis: reaction to frustration, ego-defense, and hypnotizability. II. Thematic apperception approach. *Character & Pers.*, 1942, 11, 150-165.
168. SCHWARTZ, L. A. Social situation pictures in the psychiatric interview. *Amer. J. Orthopsychiat.*, 1932, 2, 124-132.
169. SLUTZ, M. The unique contribution of the Thematic apperception test to a developmental study. *Psychol. Bull.*, 1941, 38, 704. (Abstract.)
170. SYMONDS, P. M. Adolescent phantasy. *Psychol. Bull.*, 1941, 38, 596. (Abstract.)
171. SYMONDS, P. M. Criteria for the selection of pictures for the investigation of adolescent phantasies. *J. Abnorm. Soc. Psychol.*, 1939, 34, 271-274.
172. TOMKINS, S. S. Limits of material obtainable in the single case study by daily administration of the Thematic apperception test. *Psychol. Bull.*, 1942, 39, 490. (Abstract.)
173. WYATT, F. Formal aspects of the Thematic apperception test. *Psychol. Bull.*, 1942, 39, 491. (Abstract.)

#### Verbal Summator

174. GRINGS, W. W. The verbal summator technique and abnormal mental states. *J. Abnorm. Soc. Psychol.*, 1942, 37, 529-545.



175. SHAKOW, D. Schizophrenic and normal profiles of response to an auditory apperceptive test. *Psychol. Bull.*, 1938, 35, 647. (Abstract.)
176. SHAKOW, D. & ROSENZWEIG, S. The use of the Tautophone (verbal summator) as an auditory apperceptive test for the study of personality. *Character & Pers.*, 1940, 8, 216-226.
177. SKINNER, B. F. The verbal summator and a method for the study of latent speech. *J. Psychol.*, 1936, 2, 71-107.
178. TRUSSEL, M. The diagnostic value of the verbal summator. *J. Psychol.*, 1939, 34, 533-538.

### *Miscellaneous Projective Techniques*

179. BENDER, LAURETTA. A visual motor Gestalt test and its clinical use. *Res. Monogr. Amer. Orthopsychiat. Ass.*, 1938, No. 3.
180. BUHLER, CHARLOTTE. The ball and field test as a help in the diagnosis of emotional difficulties. *Character & Pers.*, 1938, 6, 257-273.
181. BUHLER, CHARLOTTE & KELLEY, G. *The world test*. New York: Psychological Corporation, 1941.
182. DESPERT, J. LOUISE. Technical approaches used in the study and treatment of emotional problems in children. Part II. Using a knife under certain conditions. *Psychiat. Quart.*, 1937, 11, 111-130.
183. DESPERT, J. LOUISE. Technical approaches used in the study and treatment of emotional problems in children. Part I. The story: a form of directed phantasy. *Psychiat. Quart.*, 1936, 10, 619-638.
184. DUBIN, S. S. Verbal attitude scores predicted from responses in a projective technique. *Sociometry*, 1940, 3, 24-48.
185. FITE, MARY D. Aggressive behavior in young children and children's attitudes toward aggression. *Genet. Psychol. Monogr.*, 1940, 22, 151-319.
186. FOULDS, G. The child's response to fictional characters and its relationship to personality traits. *Character & Pers.*, 1942, 10, 289-295.
187. HAGGARD, E. A. A projective technique using comic strip characters. *Character & Pers.*, 1942, 10, 289-295.
188. HAGGARD, E. A. & SARGENT, HELEN. Use of comic strip characters in diagnosis and therapy. *Psychol. Bull.*, 1941, 38, 714. (Abstract.)
189. HOROWICZ, RUTH E. A pictorial method for the study of self-identification in preschool children. *J. Genet. Psychol.*, 1943, 63, 135-148.
190. HOROWICZ, RUTH E. Racial aspects of self-identification in nursery school children. *J. Psychol.*, 1939, 7, 91-101.
191. HOROWICZ, RUTH & HOROWICZ, E. H. Development of social attitudes in children. *Sociometry*, 1938, 1, 301-338.
192. KELLY, G. A. & BISHOP, F. A projective method of personality investigation. *Psychol. Bull.*, 1942, 39, 599. (Abstract.)
193. KERR, MADELINE. The validity of the mosaic test. *Amer. J. Orthopsychiat.*, 1939, 9, 232-236.

194. LOWENFELD, MARGARET. The world pictures of children. A method of recording and studying them. *Brit. J. Med. Psychol.*, 1939, 18, 65-100.
195. MARQUIDT, SYBIL. A technique of inquiry into individual personality. *Psychol. Bull.*, 1941, 38, 598. (Abstract.)
196. MIRA, E. Myokinetic psychodiagnosis: a new technique for exploring the conative trends of personality. *Proc. R. Soc. Med.*, 1940, 33, 9-30.
197. MURPHY, LOIS B. *Social behavior and child personality*. New York: Columbia Univ. Press, 1937.
198. PICKFORD, R. W. Imagination and the nonsense syllable: a clinical approach. *Character & Pers.*, 1938, 7, 19-40.
199. PROSHANSKY, H. M. A projective method for the study of attitudes. *J. Abnorm. & Soc. Psychol.*, 1943, 38, 393-395.
200. RAPAPORT, D. The Szondi test. *Bull. Menninger Clin.*, 1941, 5, 33-39.
201. SARGENT, HELEN. An experimental application of projective principles in a paper and pencil personality test. *Psychol. Monogr.*, 1944, 57, No. 5, 1-57.
202. STERN, W. Cloud pictures: a new method for testing imagination. *Character & Pers.*, 1938, 6, 132-146.
203. TUDDENHEIM, R. The reputation test as a projective technique. *Psychol. Bull.*, 1931, 38, 749. (Abstract.)
204. WERTHAM, F. & GOLDEN, LILI. A differential diagnostic method of interpreting mosaic and colored block designs. *Amer. J. Orthopsychiat.*, 1941, 98, 124-131.

### Films

205. FISHER, M. S., STONE, L. J., & BUCHER, J. Balloons: demonstration of a projective technique for the study of aggression and destruction in young children. New York: New York Univ. Film Library, 1941. (650 ft., sound.)
206. FISHER, M. S., STONE, L. J., & BUCHER, J. I. Blocking games. II. Frustration game. New York: New York Univ. Film Library, 1942. (1200 ft., sound.)
207. FISHER, M. S., STONE, L. J., & BUCHER, J. Finger painting: children's use of plastic materials. New York: New York Univ. Film Library, 1941. (790 ft., Kodachrome.)

### SECTION III

### *Supplementary References*

208. ALLPORT, F. H. Teleonomic description in the study of personality. *Character & Pers.*, 1937, 5, 202-214.
209. ALLPORT, G. W. Personal documents in psychological science. *Soc. Sci. Res. Coun. Monogr.*, 1942, No. 49.
210. ALLPORT, G. W. *Personality: a psychological interpretation*. New York: Holt, 1937.
211. ALLPORT, G. W. The psychologist's frame of reference. *Psychol. Bull.*, 1940, 37, 1-28.



212. BALDWIN, A. L. The statistical analysis of the structure of a single personality. *Psychol. Bull.*, 1940, 37, 518-519. Abstract. (See also discussion in 209.)
213. BALKEN, EVA RUTH. Psychological researches in schizophrenic language and thought. *J. Psychol.*, 1943, 16, 153-176.
214. BARTLETT, F. C. An experimental study of some problems of perceiving and imagining. *Brit. J. Psychol.*, 1916, 8, 222-266.
215. BENJAMIN, A. C. *An introduction to the philosophy of science*. New York: Macmillan, 1937.
216. BEERS, C. *A mind that found itself*. New York: Longmans, Green, 1908.
217. BINET, A. & SIMON, T. The development of the intelligence in children. *L'Année Psychologique*, 1905, 11, 163-244.
218. BLEULER, E. Upon the significance of association experiments. In Jung, C. G., *Studies in Word Association*. (Trans. by M. R. Eder.) New York: Moffat Yard, 1919. Pp. 107.
219. BOYNTON, P. L. & WALSWORTH, B. M. Emotionality test scores for delinquent and non-delinquent girls. *J. Abnorm. Soc. Psychol.*, 1943, 38, 87-93.
220. BRIDMAN, P. W. *The nature of physical theory*. Princeton: Princeton Univ. Press, 1936.
221. BRITTAİN, H. L. A study of imagination. *Ped. Sem.*, 1907, 14, 137-207.
222. BROWN, J. F. *Psychodynamics of abnormal behavior*. New York: McGraw-Hill, 1940.
223. CARTWRIGHT, D. & FRENCH, J. R. R., JR. The reliability of life history studies. *Character & Pers.*, 1939, 8, 110-119.
224. CLARK, L. P. The phantasy method of analyzing narcissistic neuroses. *Med. J. and Rec.*, 1926, 123, 154-158.
225. CONKLIN, E. S. The foster child phantasy. *Amer. J. Psychol.*, 1920, 32, 59-76.
226. CROSSLAND, H. R. The psychological meethod of word-association. *Univ. Oregon Psychol. Ser.*, 1929.
227. DAVIS, F. P., JR. Diagnostic methods in clinical psychology. Unpublished doctoral dissertation. Univ. Texas, 1943.
228. DUFF, I. F. A psychoanalytic study of a fantasy of St. Thérèse de l'enfant Jésus. *Brit. J. Med. Psychol.*, 1926, 5, 345-353.
229. FREUD, ANNA. *The ego and the mechanisms of defense*. (Trans. by C. Baines.) London: Hogarth, 1937.
230. FREUD, S. *General introduction to psychoanalysis*. New York: Boni & Live-right, 1920.
231. GALTON, F. *Inquiries into human faculty and its development*. London: J. M. Dent Co., 1883.
232. GARRETT, H. E. & ZUBIN, J. The analysis of variance in psychological research. *Psychol. Bull.*, 1943, 40, 233-267.
233. GRANT, D. A. On "The analysis of variance in psychological research." *Psychol. Bull.*, 1944, 41, 158-166.
234. GREEN, G. H. *The daydream. A study in development*. London: Univ. London Press, 1923.

235. GRIFFITHS, R. A. *A study of imagination in early childhood, and its function in early development*. London: Kegan Paul, 1935.
236. HALL, G. S. Notes on cloud fancies. *Ped. Sem.*, 1903, 10, 96-100.
237. HARVEY, N. A. Imaginary playmates and other mental phenomena of children. Ypsilanti, Mich.: State Normal College, 1918.
238. HULL, C. L. & LUGOFF, L. S. Complex signs in diagnostic free association. *J. Exp. Psychol.*, 1921, 4, 111-136.
239. JOHNSON, W. The quantitative study of language behavior. *Psychol. Bull.*, 1931, 38, 528.
240. JOHNSON, W., FAIRBANKS, HELEN, MARY B., & CHOTLOS, J. W. Studies in language behavior. *Psychol. Monogr.*, 1944, 56, No. 2.
241. JOHNSON, W. Language and speech hygiene. Chicago: Institute of General Semantics, 1939.
242. JUNG, C. G. The association method. *Amer. J. Psychol.*, 1910, 21, 219-269.
243. JUNG, C. G. *Studies in word association*. (Trans. by M. D. Eder.) New York: Moffatt Yard & Co., 1919.
244. KANTOR, J. H. Current trends in psychological theory. *Psychol. Bull.*, 1941, 38, 29-61.
245. KENT, GRACE H. & ROSANOFF, A. The study of association in insanity. *Amer. J. Insan.*, 1910, 67, 37-96.
246. KORSYBSKI, A. *Science and sanity*. Lancaster: Science Press, 1933.
247. LEHRMANN, P. R. Phantasy in neurotic behavior. *Med. J. Rec.*, 1927, 126, 342-344.
248. LEHRMANN, P. R. The phantasy of not belonging to one's family. *Arch. Neurol. Psychiat.*, Chicago, 1927, 18, 1015-1025.
249. LEWIN, K. *A dynamic theory of personality*. New York: McGraw-Hill, 1935.
250. LINQUIST, E. F. *Statistical analysis in educational research*. New York: Houghton Mifflin, 1940.
251. LLOYD, WILMA. Some aspects of language as significant of personality. *Psychol. Bull.*, 1941, 38, 747. (Abstract.)
252. MACCUDY, J. Phantasy of the mother's body in the Hephaestus myth and a novel by Bulwer-Lytton. *Psychoanal. Rev.*, 1920, 7, 295. (Abstract.)
253. MASLOW, A. H. Dynamics of personality organization. *Psychol. Rev.*, 1943, 50, 514-539, 541-558.
254. MURPHY, G., MURPHY, LOIS, & NEWCOMB, T. *Experimental social psychology*. New York: Harper, 1938. Pp. 279-299.
255. NEWMAN, S. Personal symbolism in language patterns. *Psychiatry*, 1939, 2, 177-183.
256. NEWMAN, S. & MATHER, V. G. Analysis of spoken language of patients with affective disorders. *Amer. J. Psychiat.*, 1938, 94, 913-942.
257. PIAGET, J. *Judgment and reasoning in the child*. New York: Harcourt, Brace, 1928.
258. PIAGET, J. *Language and thought in the child*. New York: Harcourt, Brace, 1932.



259. ROBINSON, E. E. The compensatory function of make-believe play. *Psychol. Rev.*, 1920, 27, 434-438.
260. ROSANOFF, A. J. *The free association test*. (Reprinted from *Manual of Psychiatry*.) New York: Wiley, 1927.
261. SANFORD, F. H. Speech and personality. *Psychol. Bull.*, 1942, 39, 811-845.
262. SEARS, R. R. Survey of objective studies of psychoanalytic concepts. *Soc. Sci. Res. Coun. Monogr.*, 1943, Bull. 51.
263. SEASHORE, R. H. *Fields of Psychology: An experimental approach*. (Chap. 40, Convergent trends in psychological theory.) New York: Holt, 1942.
264. SNEDECOR, G. W. Analysis of variance and covariance of statistically controlled grades. *J. Amer. Statist. Ass.*, 1935, 30, 263-268.
265. SOUTHARD, E. E. On the application of grammatical categories to the analysis of delusions. *Phil. Rev.*, N.Y., 1916, 25, 424-455.
266. STOUFFER, S. A. An experimental comparison of statistical and case history methods of attitude research. (Unpublished.) Chicago Univ. Library, 1930. (See discussion in 209.)
267. SYMONDS, P. M. *Diagnosing personality and conduct*. New York: Century, 1931.
268. THURSTONE, L. L. The stimulus response fallacy. *Psychol. Rev.*, 1922, 30, No. 5.
269. VARENDONCK, J. *The psychology of daydreams*. London: Allen & Unwin, 1921.
270. VERNON, M. D. The relation of cognition and phantasy in children. *Brit. J. Psychol.*, 1940, 31, 1-19.
271. VERNON, P. E. The matching method applied to investigations of personality. *Psychol. Bull.*, 1936, 33, 149-177.
272. VIGOTSKY, L. L. Thought and speech. (Trans. by Drs. Helen Kogan, Eugenia Hanfmann, and Jacob Kasanin.) *Psychiatry*, 1939, 8, 29-54.
273. WELLS, F. L. & WOODWORTH, R. S. Association tests. *Psychol. Rev. Mongr. Sup.*, 1911, No. 57.
274. WOODROW, H. & LOWELL, F. Children's association frequency tables. *Psychol. Rev. Mongr. Sup.*, 1916, No. 97.

*Edith Lord*

*EXPERIMENTALLY INDUCED  
VARIATIONS IN RORSCHACH  
PERFORMANCE*

*I. Introduction*

**D**URING THE past decade many new personality tests have been devised, little-used old tests have been revised, popular tests have been polished, refined, and expanded. College catalogues throughout the United States have lengthened lists of courses offered in psychology to include instruction in the administration and interpretation of personality tests in general, of "projective techniques" in particular, and of the Rorschach Ink Blot Test specifically.

Literally hundreds of research reports and observational articles have been published on the Rorschach and its clinical use. The bulk of this literature has been concerned with Rorschach patterns as they relate to psychiatric syndromes, to social or cultural groups, to successful and unsuccessful students, workers, or military groups. The practical or empirical success of the instrument has been widely reported. The operational results

Reprinted from *Psychol. Monogr.*, 1950, 60 (10, Whole No. 316) by permission of the American Psychological Association and the author.



of the Rorschach as a diagnostic instrument have been repeatedly acclaimed.

There is, however, a dearth of laboratory investigation of this important and widely-used clinical tool. In fact, there is even a question as to whether the test actually is a projective technique. Schachtel (31) offers an excellent critique of the projection hypothesis of the Rorschach test, and Bellak (3) states flatly that Rorschach's test is *not* a projective method. More clarification of these philosophical and theoretical areas underlying the Rorschach is definitely indicated.

Too, the paucity of statistical studies of reliability and validity, or of adequate norms, plus the existence of contradictory reports on these crucial points has given rise to a diapason of criticism. This cacophony has been particularly painful to the ears of the conscientious clinical psychologist who daily uses the Rorschach for diagnosis, for recommendations for therapy, for vocational or educational guidance and counseling, etc.

As Hsü (16) points out, there are two popular and disparate attitudes toward the Rorschach held by equally psychologically sophisticated persons. One maintains that the prime importance of the instrument is its revelation in symbolic terms of the personality of the individual, considered as a unique universe; the other stoutly maintains that this test, like all other psychological measuring devices, should meet at least minimum standards of statistical reliability and validity.

Hunt (17) criticizes the overemphasis in clinical psychology on quantitative data rather than qualitative, and bases his hopes for the future progress of clinical testing on an increased attention to qualitative behavior of the person tested rather than on quantitative results of the test.

Qualitatively, does a subject, in fact, satisfy one of the basic assumptions underlying all projective techniques—that his concepts reveal his consistent way of organizing experience, as measured by the ideas and feelings he projects into “meaning-free” or ambiguous stimuli? Is a subject reacting only to the “unstructured” or “semi-structured” ink blots on the Rorschach test, or is he reacting to the total perceptual and social field, which includes the administrator of the test and the affective tone of the administrative situation? If these latter two variables are altered, will the subject's perception of the ink blots vary significantly? Can a subject conceal his way of organizing experience or stimulate a pattern not his own? Can an administrator induce an altered Rorschach pattern by explicit or implicit suggestion?

There have been some very interesting explorations of these last two questions. Fosberg (11) studied attempts to fake Rorschach results, and also (12) examined protocols obtained under varied instructions. He demon-

strated that psychologically sophisticated subjects, given a neutral or control Rorschach, could not conceal their basic test patterns on two subsequent tests in which they tried to give the best, then the worst, possible impression.

Schachtel (30) emphasizes the importance of a subject's personal definition of the test situation and its demands. He also strongly presents his personal conviction that the three most common subjective definitions of the Rorschach situation are as authoritarian, competitive, or resistance situations. These definitions, he maintains, influence Rorschach results.

Levine (22, 23) reports a study of a series of protocols produced by the same subject while experiencing different situations hypnotically induced. There were marked changes in the Rorschach records of the various situations, each taking on the coloring considered characteristic of a person experiencing the situation created by hypnotic suggestion.

Luchins (27) studied the influence on Rorschach responses of situational and attitudinal factors. Among his findings was the marked need for the development of methods which would prepare a subject for the Rorschach test. A significant factor in the Rorschach results he analyzed proved to be misunderstandings of directions and subjects' hazy concepts of what was expected of them in the testing situation.

Wilkins and Adams (33) studied results of Rorschachs obtained under hypnosis, under sodium amytal, and in the normal, waking state. They found that both the chemically and psychologically induced abnormal states contributed to increased productivity on the Rorschach among subjects described as "overly cautious" or "fearful."

An interesting variation in the sort of study conducted by Levine is that of Bergman, Graham, and Leavitt (4) who reported the Rorschach records of a hypnotically regressed patient, diagnosed as a conversion hysteric. They found that the protocols varied with the level of hypnotic regression but closely followed the expectancies growing out of other clinical data.

Cofer (9) studying the changes in responses given by a patient with organic involvement while under hyoscine, found that some of the cards are more sensitive than others to situational variations.

A recently published research project of the Army Air Forces (32) included an analysis of the influence of the examiner on the number of responses obtained on Rorschach records. Of 36 *t*-ratios between the means of nine examiners, 12 were significant at the 1 per cent level of confidence and three at the 5 per cent level. Clearly this one scoring symbol (R) would seem to vary with the examiner.

Keeping this last finding in mind, and hypothesizing a continuum between the findings of Levine and Fosberg, one wonders whether the non-



hypnotized subject ever experiences attitudes toward the administrator or mood reactions to the total administrative situation which are adequate to alter the determinants of his Rorschach record significantly.<sup>1</sup> This, in brief, is the problem which led to the development of the present study.

## *II. The Problem*

**Statement of the problem:** This study of the influence of neagive and positive rapport conditions on Rorschach performance is based on the hypothesis that the perception of ambiguous or "unstructured" stimuli is not influenced by the perceiver's affective reaction toward the total stimulus field, which includes the administrator—the person presenting the "unstructured" stimuli. If an experimental test proves this hypothesis false, implications are legion for the interpretation of test results and for the training of administrators.

Of importance is the possibility that this study—a test of the foregoing hypothesis—may reveal whether or not the administrator is a significant part of the stimulus field during the administration of projective tests.<sup>2</sup> The theory behind projective testing is that the subject structures stimuli presented to him in nonstructured or semi-structured form. If a subject is projecting meanings, ideas, and feelings into ambiguous stimuli uninfluenced by the person presenting the stimuli or by the affective tone of the administrative situation, there should be no significant variation in his structuring of the stimuli if either the administrator or the affective tone of the administration is varied.

The problem of transference arises in all therapies, regardless of type, and the findings of this project might conceivably throw some needed light on the subject of transference elements. If the requirements for positive transference could be determined or the situations in which it flourishes could be measured, the selection and training of therapists could be more clearly defined. It is thought that the present study may hold implications for this related problem.

Outside the realm of clinical psychology, this study may have implications for social psychology. Too, the manner in which interpersonal rela-

1. An explanation of the Rorschach scoring symbols used in this study is given in the Appendix. (See p. 141.)

2. Cognizance is taken here of the fact that one cannot amplify specific findings related to the Rorschach ink blots into generalizations concerning projective techniques as a category; however, it is thought that any light on the variability or stability of one projective technique may, at least, suggest areas of desirable investigation for all other clinical tools falling within the "projective" category.

tionships affect perceptual processes is of great current interest to systematic psychology (14) as well as to the field of clinical psychology. While the present study does not answer the questions here posed, it does provide implications for future research in these areas.

Within the framework of the therapeutic situation, there may be at work individual variables which cause one subject to function more "normally" in a threatening environment, another to function more productively or adequately in an environment of heightened acceptance or permissiveness. In so far as the therapist's personality may structure these mood-situations, the results of this project could be revealing.

Designing an experiment to test whether Rorschach records will vary with experimental variation of administrator and tone of administration, one must ask himself, from a theoretical point of view, what variations might be expected. Remembering the experiments of Levine, one might hypothesize radical shifts in protocols from one administration to another if the administrations were adequately loaded with different affective tones. Remembering the experiments of Fosberg, wherein psychologically sophisticated subjects proved incapable of successfully simulating the Rorschach records of personality types other than their own, one might predict no significant variation in total protocol, regardless of administrator or of type of administration.

Somewhere between these extremes, one may ask what sort of variation might be expected (a) if a group of subjects were given control Rorschach administrations, later were given positively loaded<sup>3</sup> Rorschach administrations, and still later negatively loaded administrations; (b) if a group of subjects were given Rorschachs by three different administrators at intervals of four to six weeks. Following are the hypotheses which seem to follow from the cited relevant studies, and which seem reasonable from the standpoint of personality dynamics and the Rorschach symbols or projective response-categories which are widely believed to be related to these dynamics.

1. There will be a variation in the number of responses elicited by the various administrators, regardless of affective variable or sequence of administration. This hypothesis is based on the A.A.F. study, cited above.

2. There will be no consistent variation in number of responses between the first, second, and third administrations; i.e., no increase or decrease in number of responses with successive administrations. However, the variation in number of responses will be related to a subject's initial performance. If a subject gives thirty or fewer responses on his first test, he will increase the

3. The word *loaded* is here used to mean *weighted* or *slanted*. See IV, C for complete exposition of this term as herein employed.



number with subsequent testings; if he gives seventy or more on the first, he will decrease his responses on subsequent testing.<sup>4</sup>

3. Even among "normal" subjects, the alterations attendant upon variations of affective tone of administration will be largely dependent upon individual differences, the more stable personality records showing less change, the more labile records showing more change. It is predicted that approximately a third of the subjects' records will show no differences or only minimal changes regardless of administrator, order of administration, or experimentally varied affective tone of administration; a third will be extremely variable, fluctuating with administration order, administrator, and tone of administration. This stability or lability will be reflected in the constancy or shift of the  $M : \text{sum}C$  ratio.<sup>5</sup>

4. Regardless of the order of administration or the affective loading, there will be some variation in the records obtained by the three administrators which will be a function of the different personalities of the administrators. Since this factor is an unknown variable, it is not possible to predict the anticipated variations specifically. This hypothesis is a logical outgrowth of the known variation of the number of responses with examiner differences (32). Responses are scored as to location areas and determinants. An increase in responses insures an increase in one or another of these scoring categories. Furthermore, since number of responses has been shown to vary with examiner differences, it is not unreasonable to hypothesize that other scoring categories of the Rorschach may be equally sensitive to administrator differences.

5. Regardless of the administrator or the sequence of administration, there will be variations in the protocols produced under the three different administrative situations: neutral or control, positive affective loading, negative affective loading. Specific variations anticipated are the following:

a. Negatively loaded administration.

1) Greater scattering of response determinants among the signs  $k$ ,  $K$ ,  $Fc$ , and  $C'$ , indicating a withdrawal from the painful situation ( $k$  or  $K$ ), a depression of a situational sort ( $C'$ ), and a sensitive approach to the total stimulus environment ( $Fc$ ) as a way of handling anxiety provoked by the situation. These signs will tend to increase at the cost of the movement and color responses.  $Fm$  will appear with greater frequency as an indicator of tension in the face of a threatening situation.

2) There will be an increase in white-space responses ( $S$ ) as an indication of negativistic or oppositional tendencies.

3) There will be an increase in  $F\%$  and in  $FK + F + Fc/R$  as a reflection of an

4. Since thirty to seventy responses is usually considered the normal range, this hypothesis is based on an anticipation of the operation of central tendency.

5. The particular choice of thirds in this hypothesis is based on verbal prediction made by Dr. Bruno Klopfer.

increased effort to master the situation through intellectual control or through this control supported by heightened sensitivity and introspection.

b. Positively loaded administration.

1) There will be a greater emotional expression, or emotional relationship with the environment in this more permissive situation; this will be reflected in an increase in sumC.

2) There will be an increase in movement responses, M and FM, indicating a wider play of creativity and a greater freedom of expression of "Id drives."

### *III. Procedure*

An experimental design was constructed which aimed toward controlling as many variables as possible. The basic unit of the study was 36 males between the ages of 19 and 27. They were selected at random from a list of approximately 200 college sophomores enrolled in an introductory course in general psychology. Each subject was given three Rorschachs at four to six weeks' intervals. Each of the three Rorschachs was administered by a different person.

The three administrators, A, B, and C, were females, not different in any grossly apparent way; i.e., no one of the three was outstandingly fat or thin, ugly or beautiful. Administrators A and C were brunettes, B was a reddish-blonde. The age range among the three was twelve years, with B the youngest and C the oldest. Each administrator holds a degree in psychology and has had several years' experience in the use of the Rorschach according to the Klopfer technique.

The Rorschach tests were given in a systematic or rotated order so that twelve subjects received control tests first, twelve second, and twelve third. Twelve received negative tests and twelve received positive tests first. The second and third testing sessions were similarly varied.

No subject was tested twice by the same person. Twelve were tested first by Administrator A, second by B, and third by C. The other administrations were distributed similarly. Each administrator used exactly the same technique for the administration of the Rorschach Test, regardless of the affective loading of the pretest situation and regardless of whether the administration was the first, second, or third for the subject.

The negative and positive situations included two simple card-sorting tests administered prior to the Rorschach test for the purpose of setting the affective tone of the session. Identical directions, terminology, and techniques were used in both the negative and positive pre-tests; however, the tone of the administrations was differently slanted. Specific directions for the



administration of the three sessions were followed explicitly and without variation by the three administrators. These directions were as follows:

#### ADMINISTRATION

*Neutral.* Standard Rorschach procedure will be used. The administrator will be courteous but businesslike in manner. She must attempt to avoid either negative or positive affective loading of the situation. There will be no pre-tests nor any gathering of biographical data.

*Negative administration.* The administrator will assume the role of a harsh, rejecting, authoritarian figure. She must be deliberately unconcerned about the subject, not look at him while asking questions, preparing tests, or giving directions; never smile, give directions in a voice of dictatorial harshness, make every "Hm!" sound like a sneer.

*Positive administration.* The administrator will be personally warm, charming, appreciative in manner. She must look at subject with a smile while asking questions, preparing tests, or giving directions in an encouraging tone of voice, making every "Hm!" sound like a compliment for work well done.

The following administration will be used verbatim for *both* negative and positive administrations:

Are you Mr. Blank? Come with me! This room. Sit here.

What is your full name? You have an address? Freshman, Sophomore, Junior? Your birthdate?

Are you familiar with an ordinary deck of playing cards? (Place deck face down, with one black nine on the bottom of the pack.) When I say GO turn the cards, one at a time, and pull out all the black nines and the red tens. Put the black nines here and the red tens here (indicate areas with gestures). Work as fast as you can. Ready? Go! (Answer any questions only by repeating words in the foregoing instructions. Note any cheating, i.e., number of times subject turns two or more cards at a time; also note any errors in carrying out instructions or any requests for repetition of instructions.) Record time.

Hm! Some speed! (Leave nines and tens; shuffle remaining cards; replace, face down, before subject.)

Now, when I say GO, turn the cards, one at a time, and see how many face cards—Jacks, Queens, and Kings—you can pull out before I say STOP. Put the red face cards on the black nines and the black face cards on the red tens. Ready? Go! (Allow thirty seconds.)

STOP! (Count face cards and record number.)

Hm! Not bad!

(The Rorschach Test will follow using the same instructions as for the neutral administration.)

As the experimental design reveals, each subject follows a unique pattern. Consequently, if a subject were lost at any stage of the experiment, a substitute would have to be started from the very beginning. In order to insure completion of the design, the entire pattern was started in duplicate; i.e., seventy-two subjects were given first Rorschachs. Despite this precaution, there were several cases wherein both of the pair were lost, and a third subject (with his parallel substitute) was started through the design.

The principal cause for loss appeared to be the subject's reactions to the negatively administered session. Following this session, the subjects seemed more difficult to contact for testing appointments and frequently broke appointments or failed to appear as scheduled. Thirteen of the original seventy-two subjects were dropped following the negatively administered test because they had failed to keep from three to five appointments for the subsequent test. Two subjects were dropped for the same reason following positively administered sessions. Eight of the thirteen were given negatively loaded tests during the first session, five during the second. The losses, by examiner, were approximately equal, two losing four subjects each, the third losing five after negative administrations. The only reason for dropping these subjects, rather than persisting until an appointment was kept, was the time factor. If a period longer than six weeks elapsed between sessions, the subject was dropped and his parallel substituted in the pattern.

With this evidence in mind, one may speculate that the differences between negative results and positive or neutral results might have been greater if methodology had permitted retention of the negative records of the subjects who apparently rejected the entire project after experiencing the negative administration.

A few subjects were lost for other reasons. Three moved from the community before completing the three sessions; two were ill between sessions and unable to meet the requirement of a maximum time-lapse of six weeks between tests.

In every other case, testing of parallel partners was dropped as soon as the principal subject had completed his pattern and thereby precluded the need for a substitute.

As Rorschach records were accumulated, they were coded and mixed. The 108 records were then scored without reference to subject, name of administrator, type or order of administration. Only main responses and determinants were used in the analyses of the records for two reasons: (1) incorporation of additional responses and determinants would unprofitably over-complicate the study; (2) there are marked differences among exponents of the Rorschach method (e.g., Beck, Buhler, and Klopfer) as to the most desirable use to make of these additional determinants in scoring and in interpretation.

#### *IV. Results of the Experiment*

In addition to the loss of subjects following negative administrations, there is other evidence that the negative situation was actually traumatic to the subjects. Although the very simple card-sorting pre-tests were designed



merely to set the administrative tone of the administration for the Rorschach, behavior on the pre-tests suggests that in many cases the subjects were reacting differently by the time the pre-tests were introduced. The initial greeting, without a smile or a direct look, plus the manner and tone of voice in which the four personal-data questions were asked, apparently caused the subjects some concern.

#### A. BEHAVIOR ON PRE-TESTS

Not a single subject on the positively-loaded administrations misunderstood the pre-test directions or made an error in carrying them out. Twelve of the thirty-six subjects made from one to six errors on the negative administrations, with a total of twenty-five errors. Failure to understand directions was counted as an error if the subject requested repetition or clarification. If he demonstrated his confusion through incorrectly carrying out the directions, he was not corrected, but the error was noted.

During the Rorschach administration, the writer recorded every audible sigh or audible laugh. The laughs and sighs elicited by this one administrator give further qualitative evidence that the subjects were, in fact, reacting differently to the different affective loadings of the administrations: Eleven of the thirty-six subjects sighed aloud from one to seven times (total of thirty-two sighs) during the free-association period of the Rorschach, i.e., during the initial presentation of the plates. One subject sighed aloud during the positive, two during the neutral, and eight during the negative administrations. Seven subjects laughed aloud once or twice with a total of nine audible laughs; three laughed during the positive, two during the negative, and two during the neutral administrations. No effort was made to tabulate smiles and merely visible sighs.

One observes that audible sighs occur with greater frequency during a cold or rejecting situation than during a warm, permissive situation or a neutral situation, but that laughing aloud occurs with approximately equal frequency during the three different situations.

#### B. VARIATION IN NUMBER OF RESPONSES

The hypothesis concerning successive increase or decrease of number of responses as a function of the number given on the initial test held well for two cases which initially gave more than seventy responses but failed to hold consistently for twenty-six cases giving fewer than thirty responses at the outset.

There is an apparent bias in testing response limits exceeded by such an uneven number of subjects as exceeded our limits; i.e., two versus twenty-

six. The explanation of this procedure rests in the earlier hypothesis based on erroneous original expectancy as to number of responses. Rorschach literature quite generally states that thirty responses can be considered the minimum expectancy with normal, intelligent, adult subjects, seventy responses the maximum expectancy. The failure of our subjects to conform to this expectancy is discussed later under the subheading, "Number of responses."

### C. ERLEBNISTYP

Hermann Rorschach's theory of personality rests on the foundation of what he called *Erlebnistyp*, a word which has been variously translated but which is consistently rendered *Experience Type* in the English translation of Rorschach's original paper. He (29) makes reference to the problem, so labeled, throughout his *Psychodiagnostik*, and insists that an individual's protocol should be interpreted within the framework of his *Erlebnistyp*, that is, his introversive, extratensive, or ambiversive orientation, his "experience-balance" in responding to stimuli from within and/or stimuli from without. Rorschach found that *Erlebnistyp* was reflected in the ratio between movement and color responses ( $M : \text{sum}C$ ).

The scoring symbol M denotes, according to Rorschach, "Form perceptions plus kinaesthetic factors"; (29, p. 25) and he adds that "The more kinaesthesia, the more stable the affectivity" (29, p. 26). "Color responses," he continues, "have proved to be the representative of the affectivity and the rule is, the more color in the test, the greater the emotional instability of the subject" (29, p. 76).

A person with more M than C in his record would be designated the M-type (introversive), described by Rorschach (29, p. 81) as having the following characteristics:

1. Predominance of personalized productivity.
2. Intensive rapport.
3. Stable affect and motility, awkwardness, insufficient adaptability to reality and insufficient extensive rapport.

A subject with more C than M would be designated the C-type, more extratensive than introversive. The general characteristics of this type, Rorschach (29, p. 83) lists as follows:

1. The urge to live in the world outside oneself.
2. Restless motility.
3. Unstable affective reactions.



Klopfer (21) has pointed out the necessity of taking into consideration other factors in the record before venturing more than a simple statement of the subject's "natural inclinations" as reflected in his M : sumC ratio. These additional considerations give information as to whether the individual follows these inclinations or is in a state of conflict over them.

For the purpose of the present study, we may confine ourselves to an analysis of the extent to which lability and stability of the M : sumC ratio exists within the framework of our experimentally varied testing situations.

Rorschach recognized the existence of temporary variations in number of M and C in a subject's record if the mood of the subject shifted between dejection and elation; however, he stated that despite the variation in absolute number, "the proportion between them (M and C) changes little or not at all" (29, p. 94).

In the 108 records produced by our thirty-six subjects, variation in absolute number ranged from 0 to 19 on the M side and from 0 to 13 on the sumC side. The proportion between the M and C was exceedingly shifty; and, in addition, there were numerous actual shifts of weight in the ratio from one side to the other. The number and direction of these variations in Experience Balance are summarized in Table I.

TABLE I  
Number and Direction of Shifts in Balance on M: sumC Ratios

*Frequency and Direction of Change*

Situation	M to C	C to M	A* to M	A to C	M to A	C to A	Stable	Per Cent Stable	Per Cent Unstable
I to II	1	2	0	3	4	1	25	69	31
I to III	5	4	2	3	3	0	19	53	47
II to III	6	6	1	4	1	0	18	50	50
A to B	7	2	1	2	4	2	18	50	50
A to C	8	2	3	2	1	1	19	53	47
B to C	2	4	2	4	0	0	24	67	33
0 to +	2	5	2	1	2	4	20	56	44
+ to -	4	5	3	2	2	0	20	56	44
0 to -	2	7	1	2	1	2	21	58	42
Totals	37	37	15	23	18	10	184	57	43

\* The symbol A is used to denote ambi-equality of M and sumC.

From Table I it can be seen that the greatest stability of M: sumC ratio exists between administrations I and II; even here, 31 per cent of the cases shifted from one balance to another; i.e., there was variation *within* approximately a third of the persons tested, as measured by an individual's production of more movement than color responses in one situation, more

color than movement responses in another situation. A close second in rank order of stability is that between records obtained by administrators B and C, wherein 67 per cent maintained a constant balance.

The greatest number of shifts in balance occurred equally between Administrations II and III and between Administrators A and B. In both these categories, exactly half the cases maintained a constant balance and half shifted in one direction or another.

In reading Table I, it must be remembered that a shift, from M to C for example, means that the subject actually showed more M than C under the first condition considered and more C than M under the second condition listed.

A first glance at the totals in the columns labeled "Per Cent Stable" and "Per Cent Unstable" might suggest that approximately half the subjects maintained a constant stability throughout the experiment regardless of sequence, administrator, or type of administration. Such a conclusion cannot be drawn from Table I, however, since stability therein is noted within, but not between, the variable situations.

An examination of the data reveals that only fourteen of the thirty-six cases maintained a constant balance on the M: sumC ratio throughout all variations in testing situations. Of these fourteen, ten held a balance weighted on the M side of the ratio, four showed consistent weighting on the sumC side, and none maintained ambi-equality of weighting. One may, therefore, conclude that 39 per cent of the subjects retained a stable Experience Type throughout the experiment, approximately three fourths of whom were introversively oriented, a quarter of whom were extratensively oriented.

These conclusions, then, would support the prediction stated earlier in this study that approximately a third of the subjects would produce stable records. However, on the basis of Rorschach's description of the M-type and C-type personalities, noted earlier in this discussion, one would have suggested that the fourteen subjects with stable records would *all* have been of the M-type. Table I shows us that A-type, M-type, and C-type personalities in one situation can become another personality type in another situation. The additional analysis of the data shows us that both M-type and C-type personalities are capable of maintaining their personality type in the face of varying situations, but that the odds are three to one that the subject who maintains such constancy will be an M-type personality.

#### D. STATISTICAL TREATMENT OF THE DATA

In order adequately to test the principal hypothesis of this experiment concerning the influence of negative and positive rapport conditions on Ror-



schach performance, and the numerous collateral hypotheses—the influence of successive administrations and different administrators, and the kind and amount of variation in the numerous scoring categories—one must look to statistical analysis.

To justify the greatest confidence in probabilities obtained from analysis of the data, *t*-ratios were computed for the numerous pairs of series offered by the data.<sup>6</sup> The particular *t*-ratio used was Fisher's formula as described by Lindquist (26, p. 58 ff.) under the title "The Significance of a Difference in the Means of Related Measures":

$$t = \frac{M_0 - M_H}{\frac{\text{Sum } d^2}{n(n-1)}}$$

These *t*-ratios (a total of 243) and the interpretations of them are presented in the following pages.

In working the *t*-tests, data were fractionated in the following manner: All tests given during the first session were grouped, regardless of who gave them or of what kind of affective tone marked the administration; this procedure was followed for the second and third administrations as well. All tests given by Administrator A were grouped, regardless of whether they were administered first, second, or third in the series, regardless of whether the tone of administration was positive, negative, or neutral. A similar grouping was made for tests given by Examiners B and C. All tests given during a positively loaded test situation were grouped without respect to who administered them or to the serial number of the situation. Identical groupings were made for tests administered in the negative and the neutral situations.

This combination of heterogeneous data, of course, introduces a special problem in the interpretation of the results of the *t*-tests. If variance is found due to administrations, this variance is operating when testing the differences between administrators and between types of administrations. Where significant variance exists only within one of the three major groups of variables, this problem loses some of its importance. If, however, variance is shown to exist as a function of the repetition of the test, this variance will operate to enlarge the sigma, hence enlarge the denominator, of the *t*-tests

6. Cognizance is taken here of the criticism to which this study is subject as a result of the practice of treating scoring categories as unique parts of a whole without reference to the whole of which they are an integral part. The writer can think of no other method of experimental approach to the investigation of the Rorschach and the personality factors it is judged to measure except by analysis of the separate elements which contribute to the whole. Recognizing the imperfection of the approach, the plea is offered that it remains the best of all possible existing approaches.

between other variables. The influence would produce spuriously low  $t$ -ratios among the other groups of variables.

This weakness of the  $t$ -test, which inevitably exists when data are fractionated and combined as in the present study, is not a serious drawback if one bears in mind that  $t$ -ratios approximating zero (say, under 1.000) throughout one whole set of data strongly indicate no real differences; if one remembers, also, that any significant or very significant  $t$ -ratio in one of the three groups of variables within a set will contribute to spuriously low  $t$ -ratios among the other groups, and that, therefore, a marginal  $t$ -ratio would be significant if the variance due to other factors were partialled out.

One other element of the statistical treatment demands discussion: Computation of  $t$ -ratios has been made on mean frequencies of scoring categories regardless of whether these scores are expressed as percentages of the total responses or as absolute numbers. While the percentages are not seriously affected by the variation in total number of responses, the absolute numbers are. For example, if the positive administration produces more responses than the negative, there will not necessarily be any alteration in the per cent of whole responses; there will, however, necessarily be an increase on the positive administration in the frequency of certain of the response determinants. On a statistical level, one is eager to discover *which* determinants increase; on an interpretative level, one attempts to explain *why* one determinant increases and another does not, whether the increase is a concomitant of increased responses or not.

If there were a significant difference in number of responses from the first to the second to the third administration, any differences in tabulated scoring symbols would have to be held suspect as an artifact or the frequencies would have to be reduced to ratios. However, no such differences in  $R$  were found to exist.

#### E. STATISTICAL FINDINGS

*Number of responses.* The greatest constancy in mean number of responses proved to be that between administrations. It is interesting to note here that Rorschach literature is consistent in reporting that the normal number of responses is around thirty with normal ranges reported from fifteen to seventy-five.

Hermann Rorschach: Normal subjects generally give from 15 to 30 responses, rarely less than 15, often more than 30 (29, p. 21).

Klopfer and Kelley: The range of responses found most frequently in all large-scale investigations of adults seems to be between twenty and forty (21, pp. 208-209).

Brussel and Hitch: Normal is around 34. . . . Normal range is from 25 to 75 (6, p. 3).



Bochner and Halpern: The average [number of responses] seems to fall between 20 and 50 (5, p. 71).

Charlotte Buhler gives a minus weighting for 3 for fewer than twenty-five responses (7, p. 11).

As an aside, one wonders if these published means and ranges are not a bit high. Only on the positive administration and under C's administration do our subjects reach or exceed means of thirty responses.

Further in this vein, it should be noted that the previously cited study conducted by the Army Air Forces (32) shows the average number of responses for servicemen undergoing classification testing to be only twenty. Perhaps this particular Army test situation stimulated an inhibition of responses. Perhaps, too, the special population of the present study experienced a special kind of inhibition of response.

Perhaps, on the other hand, the figure thirty is not the true mean for the population as a whole, is not the appropriate cut-off point for a normal number of responses. Possibly separate norm-groups should be established for various segments of the population.

Although there is an increase in the mean number of responses on both the negatively and positively loaded administrations over the neutral administration, with the positive eliciting more than the negative, the variability within each type of administration is large enough to keep the differences below the critical cut-off point for statistical significance higher than the 10 per cent level of confidence which exists between the positive and the neutral administrations.

There is a very significant difference, however, between the number of responses elicited by Administrators A and C. Whether giving a first, second or third administration, whether being negative, positive, or neutral, Administrator C obtained, on the average, a sufficiently larger number of responses virtually to rule out chance as a cause of the difference (1 per cent level of confidence).

*Average time per response.* Very significant differences were found in the average response time between the first test and both the second and the third tests. These differences can be explained rather easily in terms of familiarity with the task. On being first confronted with the Rorschach plates, the subjects uniformly spent more time formulating their concepts than they spent on subsequent reassociation with the stimuli. The reduction in response-time is uniform in direction from beginning to end of the series, despite the lack of a significant difference between the second and third administrations.

There is a consistent tendency toward a difference in response times between Administrator C and the other two administrators; in fact, the dif-

ference between A and C is significant at the 5 per cent level when rounded to two decimal places. Apparently the subjects, on the average, lingered longer over their responses when giving them to Administrator C.

*Crude control.*  $F\%$  is considered to be one of the key determinants in a Rorschach protocol. It is the scoring symbol used to indicate the percentage of responses determined solely by the shape, contour, or outline of the blot. Briefly, it is interpreted as a measure of the control exercised by the subject. Buhler and Lefever (8) designate it as a measure of rational functioning. Excessive  $F\%$  (50 per cent or above) is considered to be an indication of undue constriction, or an accompaniment of depression or anxiety.

While the means of the various measures of our subjects stay well under the 50 per cent upper limit for normalcy, we may speculate that significant increases in  $F\%$  mean increased depression, anxiety, or constriction. No such significant increases are found as a function of the number or the type of the administration; however, it is interesting to observe that the lowest mean is found for the positive administration and the highest for the negative.

Again we find significant differences among administrators. A differing at the 1 per cent level with C and at the 5 per cent level with B. There is no difference between B and C.

*Refined control and its components.* The  $FK + F + Fc$  per cent is another measure of control; however, it is distinguished from  $F\%$  alone by the quality of the control it reflects. Klopfer (21) has called the latter "crude control" and the former "refined control." Separately considered, the determinant  $FK$  (vista concept) is usually interpreted as reflecting introspection or insight;  $Fc$  (texture concept), as reflecting tact or sensitivity.

Although there is not a critically significant difference in the amount of refined control displayed in the various types of administration, the means show a consistent increase from the lowest mean on the positive administration to the highest mean on the negative administration, paralleling the tendency on  $F\%$  or crude control.

A significant difference between the first and the third administration suggests several alternative possibilities. The mere repetition of the Rorschach probably brought about a degree of loss of spontaneity in responding to the stimuli. This was replaced successively more frequently by increased refined control of reactions to the stimuli.

While there is practically no difference between Administrators A and B in the average percentage of responses with this combination of determinants, Administrator C differs at the 10 per cent confidence level from A and differs significantly from B.

Further light may be thrown on the factors influencing these differences



in the combination-score by inspecting the components that contribute to it.  $F\%$  has already been analyzed. The raw data show a low frequency of FK and Fc responses. The responses, however, occur with adequate frequency and with sufficient approach to normality of distribution to permit application of the  $t$ -ratio statistic.

A  $t$ -ratio for Fc between Administrators B and C was not computed since the difference between means of .05 is so small as to insure a  $t$ -ratio of approximately zero. Despite the low frequency and small differences in means between A and the other two administrators, the variability was sufficiently small to produce a significant difference between A and C, and a very significant difference between A and B. The latter two administrators consistently elicited more Fc responses than did A. This observation would tend to rule out Fc as a contributing factor to the significant difference between B and C on the  $FK + F + Fc$  percentage. Considering, too, the insignificant  $t$ -ratio between B and C on  $F\%$ , one might suppose that despite very low frequency and apparent scatter, the FK responses are the greatest single contributing factor to the  $FK + F + Fc$  per cent difference between B and C; however, this convenient supposition is not supported by the data, as will be subsequently shown.

It is interesting to note that  $F\%$  (crude control) was sufficiently higher for A than for B or C to produce significant and very significant differences respectively; whereas, Fc (tact or sensitivity) was sufficiently *lower* for A than for B and C to produce very significant and significant differences respectively. It is not surprising, then, that when these two determinants, F and Fc, are summed with a third determinant FK the differences between A and both B and C disappear but the difference between B and C is exposed.

The scoring category FK is used for concepts incorporating vista, distance, perspective. There are no significant differences in FK on the successive administrations, although the means reflect a consistent decrease in the use of this determinant. Nevertheless, no inferences can be made from such data.

The lack of significant differences among administrators also permits no interpretation, inference, or conclusion. We may only say that our analysis of FK offered no clarification of the questions growing out of the findings for  $FK + F + Fc$  per cent. This latter category, then, appears to be a unique whole, not paralleling any one of its three components in its stability or variation under the varied situations.

The affective loading of the administrations had a significant degree of influence on the scoring category FK. Both the negative and the positive administrations show more mean FK than does the control administration.

The difference between the neutral and the positive administrations is significant at the 5 per cent level of confidence.

*The symbols K and k.* Responses classified as K are interpreted as a sign of "free-floating" anxiety or of insecurity. Douglas (10) found the use of this type of response serving "as conversational material until something further can be found."

K responses remain quite stable through the successive administrations. There is a noticeable difference, however, in the amount of K elicited by the different administrators, B and C differing at the 10 per cent level of confidence.

Like K, responses classified as k are interpreted as reflecting feelings of anxiety, insecurity, or inadequacy; however, the two categories are differentiated by the control factors implicit in a k response, absent in a K response.

The low frequency of the determinant k is reflected in low means for the various measures. Neither the variation of administrator nor the type of administration affected this scoring symbol to a degree worth noticing. There is a consistent, though insignificant, reduction in the appearance of this determinant on the successive tests, each administration eliciting somewhat fewer such responses than the preceding.

*The C' determinant.* Our thirty-six subjects produced, on the average, about one and a half achromatic color responses (C') per record without variation from one administration to the next. There was variation in the C' responses with type of administration. The increase from an average of one C' on the control Rorschachs to an average of two on the positively loaded Rorschachs boosts the confidence level between these variables to 10 per cent. Administrator A elicited the lowest frequency of C' responses, differing from both B and C at the 5 per cent confidence level.

*Movement responses.* There are no statistically significant differences higher than the 10 per cent level of confidence in the number of Fm, inanimate movement, responses produced under any of the three variations of the testing situation. The constancy in number from one administration to another is so great as to preclude the necessity of computing *t*-ratios.

Mean differences, however, exist as a function of the variation in type of administration, the *t*-ratio between the control and the positive administrations exceeding the 10 per cent level of confidence that a real, not chance, difference exists.

The animal movement responses, FM, vary little or not at all on the successive administrations, and they fail to vary significantly with variation in type of administration. However, there are differences at the 5 per cent level of confidence between Administrators A and C, and at the 1 per cent level



between B and C, Administrator C eliciting the larger number of responses in both comparisons.

The frequency of responses classified as M, human movement, varied sufficiently to reach the 5 per cent level of confidence between one pair of variables and the 10 per cent level between four other pairs. Human movement is consistently interpreted as a sign of inner adjustment or equilibrium, a capacity for the absorption of emotional stimuli whether originating from within or without.

An excess of M on the positive administration over the neutral created a difference significant at the 5 per cent level of confidence, and the positive administration shows a difference from the negative at the 10 per cent level.

Administrator C provoked the highest mean frequency of M responses, differing from both A and B at the 10 per cent level of confidence. Further, the mean number of M on the third administration was sufficiently below the number on the second to give a *t*-ratio significant at the 10 per cent level.

*Emotionality.* The C-type personality was discussed earlier in this paper, and the relationship of sumC to M was defined. Not forgetting the importance of this interrelationship, let us look at the meaning of color responses per se in Rorschach records. Color is considered to be the determinant of emotionality. Lack of color is considered an indication of constriction; excessive use of color is interpreted as emotional lability or overreaction to external emotional stimuli.

The use of color varied little on successive administrations. There was an infinitesimal difference between the subjects' sum of color responses on the neutral and the negative administrations; in fact, no significant differences occurred among any of the types of administration.

The really important differences in the use of color responses occurred with the variation of examiners. Administrator A elicited significantly fewer color responses than C and very significantly fewer than B.

*Content categories.* Normally, the sum of animal and animal-detail concepts constitutes approximately 50 per cent of the content of Rorschach concepts. Our subjects linger close to this norm. An increase in A% is usually interpreted as evidence of stereotypy of thinking or a repression of intellectual activity; a decrease in A%, contrarily, is widely interpreted as less confinement to the obvious, a broader range of thinking and/or interests.

There are completely unimportant differences in A% on successive administrations of the test. There is no appreciable difference in A% from one examiner to another; although, again, the differences between Administrator A and the other two examiners show a stronger tendency toward sig-

nificance than does the difference between B and C, that between A and C closely approaching the 10 per cent level of confidence.

The percentage of A% responses on the affectively loaded administrations differs significantly, with the negative administration producing the percentage higher than that on the positive administration. Also, the neutrally conducted tests deviate from the positive with a difference at the 10 per cent level of confidence.

There is agreement among users of the Rorschach that the number of animal responses should and do exceed the number of animal-detail responses in a normal, healthy record. Therefore, it was considered advisable to break down the A% scoring symbol into its component parts in an effort to examine the circumstances under which both animal responses and animal-detail responses may vary or remain stable.

Examination reveals no significant differences in the number of whole animal responses elicited in the various situations. Even the means do not show a difference worth comment. Certainly this analysis of the number of whole animal responses does not help appreciably to clarify the differences obtained in the A%. A look at the means and critical ratios for animal details is indicated; however, one must not overlook the fact that absolute number frequencies introduce different considerations from those under consideration when one is speaking of percentages.

In all measures, the number of A responses exceeds the number of Ad responses, indicating that our subjects on the average maintain the healthier balance regardless of variation in testing situations. With both symbols there are infinitesimal differences in frequencies from one testing situation to the next. Among administrators, C elicits the highest frequency for both A and Ad responses. Among the types of administrations, the control Rorschachs show the lowest mean for both scoring categories.

Differences at the 5 per cent level of confidence exist between Administrator A and both B and C, with A gleaning significantly fewer Ad responses than B or C.

As with other scoring categories, the number of A and Ad responses by themselves are considered to have little meaning. Whereas, an increase in Ad may indicate a tendency toward a critical attitude, this interpretation of excessive Ad in a record is made only if the sum of animal-detail and human-detail responses exceeds the sum of whole animal and whole human responses. This observation calls for an analysis of the number of human and human-detail responses.

Several measurable differences in number of whole human responses exist. The frequencies are exceedingly stable from the first through the third administration. While Administrator C provoked a larger mean



number of human responses than did either A or B, only the difference with A is great enough to produce a *t*-ratio significant at the 10 per cent level of confidence.

The means of the types of administration show an increase in human concepts on both the negative and positive administrations over the number on the control administration, with the difference between the neutral and positive rising to the 10 per cent level of confidence.

Again, the frequency of whole versus part concepts of human figures is a more important consideration than the mere number of either.

A comparison of means reveals that the records of our subjects, regardless of variables, maintain a healthy preponderance of H over Hd responses.

Administrator C evoked a greater number of Hd responses than either B or A, and the mean of C's administrations differs significantly from that of A.

*Popular responses.* The means of the popular responses elicited by the various examiners are identical. There is a slight but steady increase in mean P from the first to the second to the third administration, which may easily be explained in terms of increased familiarity with the material through mere repetition.

The difference between the means of the negative and positive administrations reaches the 10 per cent level of confidence, the permissive situation stimulating the higher number.

*Use of ground as figure.* Neither repetition of administration nor affective loading of administration produced any difference in the number of white-space, S, responses on the Rorschach records. Neither are Administrators A and B essentially different in the amount of S they provoked on the protocols. Administrator C, however, differed from both A and B in eliciting a higher frequency of white-space responses, and differed from A at the 5 per cent level of confidence.

An earlier discussion of additional responses explained why they were not incorporated into the analysis of results on this study. An exception should be made, however, of the white-space scoring category. Among normal or near-normal subjects, white-space responses occur most frequently concomitant with reaction to inked areas of the plates. Arbitrary scoring methodology requires that, in such mixed concepts, the inked area be scored as the main response location and the white space be scored as an additional response location. Consequently, many clear-cut reactions to white space are buried among the additional responses. Better to understand the actual amount of and variability in the use of white space, it was decided to sum the absolute number of main and additional S responses and analyze the distribution of this total reaction to ground.

The total use of white spaces by our subjects varies significantly with type of administration and among administrators. The smallest amount of total  $S + (s)$  was elicited on the neutral or control administrations, the most on the positive administrations. While the actual mean difference between neutral and negative administrations is small, the direction of change is so consistent, the variability within the columns so small, that the difference reaches the 5 per cent level of confidence. The higher mean difference between neutral and positive administrations is accompanied by sufficiently increased variability to hold the difference down to a 10 per cent level of confidence.

Among administrators there exist real differences in the amount of sum- $S$  produced, regardless of number or type of administration. A evoked the least, and C the most, use of white space. The difference is significant at the 5 per cent level of confidence. C also exceeded B in elicitation of total  $S$  responses with a difference significant at the 10 per cent level. The most significant difference, however, is between administrators A and B, with a  $t$ -ratio at the 1 per cent level of confidence.

*Percentage of responses to all-color cards.* Color, as previously stated, is interpreted as a correlate of emotionality. Since cards VIII, IX, and X on the Rorschach test are composed exclusively of bright-color inks, they are construed to have particular significance in reflecting emotional components of a subject's personality. The normal expectancy of responses on these all-color cards is approximately 40 per cent of the total responses.

Our subjects consistently approximate normal expectancy, with little variation from one administrative situation to another. The largest difference occurs between the neutral and negative administrations, where the  $t$ -ratio rises to a 10 per cent level of confidence.

Apparently the overrejecting situations brought about a measurable decrease in response to external emotional stimulation, creating a dampening effect on the subjects.

*Manner of approach.* The number of whole responses in a Rorschach record provides two separate clues for interpretative analysis: (1) a clue as to the percentage of responses which are based on incorporation of the total blot-stimulus in the concept-formation; (2) a clue as to the ratio between whole responses and movement responses, i.e., the  $W : M$  ratio.

The first of these clues will be explored later upon the presentation of means and  $t$ -ratios for  $W\%$ . The second requires the reintroduction of the means for the scoring symbol  $M$ .

There is a steady decrease in the mean number of whole responses with each repetition of the test, the difference between number of  $W$  responses



on administrations I and III being significant at the 5 per cent level of confidence.

If the  $W : M$  ratio much exceeds  $2 : 1$ , the interpretation is usually made that the subject is overextending himself; i.e., he is striving beyond his means of achievement. It is not considered a healthy sign. Let us compare our  $W : M$  mean ratios for any evidence of this imbalance.

One ratio that was obtained on the second administration achieves the ideal relationship of  $2 : 1$ . With one exception, all other ratios show heavier  $W$  weighting; however, the ratio  $3 : 1$  is not sufficiently in excess of the ideal to merit interpretation of a significant imbalance. The ratio on the neutral administration of  $4 : 1$ , however, is twice the expectancy ratio. On the basis of Rorschach literature, one would have predicted a  $2 : 1$  ratio on this control administration, regardless of what other predictions attended the other administrations.

The really crucial interpretative factor related to whole responses is the scoring symbol  $W\%$ —the percentage of total responses which use all of the inked surface of the cards within the framework of a single concept. "In a normal record," say Brussel and Hitch, "about 25 per cent of whole answers are expected" (6). The Individual Record Blank for scoring Rorschach protocols, developed by Klopfer and Davidson, and in wide use, shows a normal range of  $W\%$  extending from 20 to 40. Every mean in the present study exceeds not only the figure given as a normal mean but also the figure given as the upper limit in the range of normal expectancy for  $W\%$ . One might assume that the stated norms are in error; however, it seems less arrogant to consider the possibility that the subjects in the present study deviate from the population as a whole somewhat.

Examination supports the earlier finding of reduction of whole responses as a mere function of repetition of the test. When working with percentages rather than absolute numbers, this reduction in tendency is more dramatically revealed. While the decrease in  $W\%$  is insignificant from the first to the second administrations, the decrease is significant at the 1 per cent level between the third and both the first and the second administrations.  $C$ 's protocols show the lowest mean percentage of whole responses and differ at the 5 per cent level of confidence with  $A$ 's records, which show the highest mean  $W\%$  among administrators.

While none of the differences in means of varied types of administration are statistically significant, it is interesting to observe that the negative administration shows the highest percentage of whole responses and the positive situation the lowest percentage.

As  $W\%$  decreases, it is mathematically certain that other location subdivisions will increase. The nature of these increases can be explored only

by analyzing the frequencies in the data of large details, small details, and tiny or unusual details as they are employed by the subjects in concept-formations.

One clear-cut difference exists between Administrators A and C. Administrator A elicited the highest percentage of W% responses and the lowest percentage of D% responses; in both scoring categories, A and C differ at the 5 per cent level of confidence.

No other variation in testing situation creates differences of statistical significance. Nevertheless, consistent tendencies are revealed by the means of the various types of administrations. The negative administration provoked the lowest mean D% and the positive administration the highest. The *t*-ratio between them barely misses reaching the 10 per cent level of confidence.

The location-scoring symbol D is used when subjects form a concept using a large, insular part of the blot, frequently perceived as a separate entity by normal subjects. Between 45 per cent and 55 per cent of total responses are expected to be D responses.

Not until the third administration did the subjects show any measurable increase in the use of small usual details in the blots. The production of concepts based on these small areas remained constant from the first to the second test but differed between the second and the third to a degree significant at the 10 per cent level of confidence.

The average intelligent subject is expected to have from 5 per cent to 15 per cent of his responses in the d% category. The third administration, then, most nearly provoked d% responses up to a minimum of expectancy for our intelligent subjects. Decreased d% is not generally interpreted as having any particular significance, and increased d% is ignored unless it exceeds 15 per cent of the total responses.

Although there is a lack of significant differences between the various administrators, Administrator C has the highest mean d% within the administrator group, just missing 10 per cent *t*-values with both A and B.

Dd + S% is a sort of wastebasket category for all responses based on blot areas not defined as W, D, or d. Such responses may include tiny details, inside or edge details, unusual combinations of large or small blot areas, and the use of ground as figure. White-space responses were analyzed earlier by absolute number because of their special qualitative meaning in Rorschach interpretation. Their tabulation in combination with unusual details is a necessary part of the process of recording percentages of responses by location areas alone.

There is complete agreement among Rorschach authorities that absence of Dd + S% is normal. There is, however, some difference of opinion as to



how high this percentage may go before the upper limit of normalcy is exceeded. The range of estimates is from 2 per cent to 10 per cent, with the consensus approaching the larger figure. In any event, the records of normal persons rarely exceed 10 per cent in this category. On all variations of the testing situation, the subjects in this study remain under this 10 per cent maximum; however, the ceiling is approached in the means of third administration and of the positively administered tests. None of the differences between means are statistically significant.

While  $Dd + S$  per cent in excess of 10 per cent is interpreted as an abnormal sign, even an excess of 5 per cent in the category is, by some authorities, considered to be evidence of a pedantic approach to the test (and, by extension, to the environment and to external stimuli, in general). All nine of our means exceed 5 per cent. The foregoing interpretation is more readily made if the higher percentage of  $Dd + S$  responses occurs in conjunction with high  $W\%$ , a fact which we have already observed to exist with our group.

*Summary of t-ratios.* This completes the statistical analysis of the data on the numerous Rorschach scoring categories. The somewhat lengthy analysis of means and  $t$ -ratios by Rorschach scoring categories seemed the most clear-cut way of presenting the data of this study. However, that procedure inevitably produces some confusion as to how many statistically significant  $t$ -ratios were found and where, how frequently Administrator A or administration + had the highest means and on which types of response. These data are presented in summary form in Tables II to IV.

Table II serves to show the administrative circumstances under which the means of each of the thirty-one Rorschach response categories, herein analyzed, held the highest, middle, or lowest numerical position within its group.

Table III shows the frequency with which each variable, within each of the administrative groups, achieved the highest, middle, and lowest means. It will be noted that the positive administration and Administrator C reveal almost identical frequency patterns. No other pairs show this marked parallel tendency.

As Table IV reveals, there were, at the 1 per cent level of confidence, a total of ten differences, over four times as many significant  $t$ -ratios as one would expect by random sampling; four of these were a function of repetition of the test, six a function of examiner differences. Administrator A differed very significantly from B on three of the scoring categories and from C on two of the measures. B and C differed only once at this level of significance.

At the 5 per cent level of confidence, one would expect twelve significant differences by random samplings; whereas, we obtained twenty-one additional differences at this level alone. Combined with the ten  $t$ -values at

TABLE II  
Rank Order of Means, According to Type of Response

Type of Response	Administration			Administrator			Type of Administration		
	I	II	III	A	B	C	+	-	0
R	2	1	3	3	2	1	1	2	3
T	1	2	3	3	2	1	1	3	2
F%	3	2	1	1	2	3	3	1	2
FK	1	2	3	2	3	1	1	2	3
K	1	2	3	2	3	1	1	2	3
Fk	1	2	3	2	3	1	1	3	2
#F	3	1	2	1	3	2	2	1	3
FC	1	3	2	3	2	1	1	2	3
CF	1	2	3	3	1.5	1.5	1	3	2
C	1	3	2	1	2	3	1	3	2
FK + F + Fc/R	3	2	1	2	1	3	3	1	2
M	2	1	3	2	3	1	1	2	3
sumC	1	3	2	3	1	2	1	3	2
FM	2	1	3	2	3	1	1	3	2
Fm	2.5	2.5	1	3	2	1	1	2	3
Fc	3	2	1	3	2	1	1	3	2
C'	1	2	3	3	2	1	1	2	3
A%	2	3	1	1	2	3	3	1	2
P	3	2	1	2.5	2.5	1	1	3	2
H	2	1	3	2	3	1	1	2	3
A	2	1	3	2	3	1	1	2	3
Hd	3	1	2	3	2	1	1	2	3
Ad	3	2	1	3	2	1	2	1	3
CR%	2	3	1	3	1	2	2	3	1
W%	1	2	3	1	2	3	3	1	2
#W	1	2	3	3	2	1	1	2	3
D%	3	2	1	3	2	1	1	3	2
d%	2	3	1	3	2	1	2	3	1
Dd + S%	3	2	1	3	2	1	1	2	3
S	1	3	2	3	2	1	1	2	3
S + (S)	1	3	2	3	2	1	1	2	3

the 1 per cent level, our total is thirty-one *t*-ratios equal to or above the value required for confidence at the 5 per cent level. This figure is almost three times chance expectancy. Successive administrations accounted for two of the significant differences, and type of administration for four. The fifteen differences among examiners point up the greater frequency between A and C than between either A and B or C and B. Three of the administrator differences were between A and B, one between B and C, but eleven



TABLE III  
Frequency of Mean Position, According to Administrative Variable

Administration	Highest Mean	Middle Mean	Lowest Mean
I	13	8½	9½
II	7	15½	8½
III	11	7	13
A	5	8½	17½
B	3½	19	8½
C	22½	3½	5
+	23	4	4
-	6	14	11
0	2	13	16

existed between A and C. Combining differences at the 1 per cent and 5 per cent levels, A differs six times with B and thirteen times with C, while B and C differ a total of twice.

Differences (5 per cent level) resulting from variation in the emotional tone of the testing situation occurred twice between the neutral and negative administrations, once between the positive and neutral, and once between the positive and negative administrations.

Clearly, the largest number of significant and very significant differences at the 1 per cent and 5 per cent levels were the result of examiner differences; and, within our group of examiners, A was the greatest contributing factor to frequency of differences. The negative and neutral administrations appear with equal frequency (three times each) as contributing factors to significant differences in type of administrations. The positive administration is a factor in two of the four differences.

Table IV summarizes the seventeen *t*-ratios at the 10 per cent level of confidence. This figure brings to forty-eight the total number of *t*-values reaching or exceeding the 10 per cent confidence level, a total exactly twice as large as the twenty-four one would expect by random sampling.

At this level of significance, 10 per cent, the variations with type of administration predominated, with nine differences tending toward significance. Two differences were a function of repetition of the test, and six

TABLE IV  
Number of Meaningful *t*-Ratios by Variables

Confidence Level	I II	II III	I III	AB	AC	BC	+-	+0	-0	Cumulative Total
1%	1	0	3	3	2	1	0	0	0	10
5%	0	1	1	3	11	1	1	1	2	31
10%	0	2	0	0	3	3	2	6	1	48

occurred with change of administrators. Administrator C was a party to all of these variations, differing three times with A and three times with B. Of the nine differences occurring with variation in the tone of the administration, the positively loaded situation differed six times with the neutral administration and twice with the negatively loaded situation. The negative and positive administrations showed differences at this level of confidence only twice, the negative and neutral only once. Summarizing, the positive administration was a factor eight times, the neutral seven times, and the negative three times.

*Summary of results by scoring categories.* A few of the Rorschach scoring categories proved exceedingly stable, showing no variation from test to retest, from examiner to examiner, from positive to negative to neutral administration. These stable categories are the following: Dd + S per cent, Fk, and number of A responses.

The majority of the Rorschach scoring categories showed measurable variability as the testing situation varied. The degree of variation, together with the administrative variable which produced the variation in each shifting category, was as follows:

a. R: There was one difference at the 1 per cent level of confidence between administrators and one difference at the 10 per cent level with variation in types of affective loading of the administrative situation.

b. T: There were two differences at the 1 per cent level in average time per response on the successive administrations and one difference at the 5 per cent level between administrations.

c. M: This scoring symbol proved to be quite variable, with one difference at the 5 per cent level and another at the 10 per cent level as a function of the varied affective loading of the testing situation; two differences occurred at the 10 per cent level with variation in administrators, and one difference at the 10 per cent level grew out of mere repetition of the test.

d. FM: Neither successive administrations nor variation in affect in the testing situation influenced FM. Change of administrators, however, produced one difference at the 1 per cent level, another at the 5 per cent level.

e. Fm: This scoring symbol was resistant to change with repetition of the test or with change of examiners. It varied only with the variation in type of administration, and then only at the 10 per cent level of confidence.

f. K. Administrator differences produced one difference in K at the 10 per cent level of confidence. No other variables in the testing situations measurably influenced this determinant.

g. FK: This response determinant also remained stable through successive tests and despite variation in examiners, but it showed a 5 per cent level of significant difference with variation in affective loading of the testing situation.

h. Fc: Examiner difference again is evident with this determinant, there being both a 1 per cent and a 5 per cent difference among administrators. Neither retesting nor affective loading affected the mean number of Fc produced.



i. C': This scoring symbol varied once at the 5 per cent level with repetition of the test, twice at the 5 per cent level with change in administrators, and once at the 10 per cent level with variation in tone of administration.

j. sumC: The composite score sumC remained stable from one test to the next and despite variation in affective loading of the testing situation; however, the mean scores varied significantly twice with change of administrators, once at the 1 per cent level, once at the 5 per cent level.

k. F%: This factor also resisted change with number and type of administration; however, it, too, varied with the examiner once at the 1 per cent level, again at the 5 per cent level.

l. A%: The difference in the type of administration produced two differences in A%, one at the 5 per cent level, the other at the 10 per cent level. Neither administrators nor successive testing influenced the means significantly.

m. P: The number of popular responses remained somewhat constant throughout the experiment. One difference, at the 10 per cent level, occurred with variation in the affective loading of the testing situation.

n.  $FK + F + F_c / R$ : The affective loading of the administration did not influence this composite scoring factor; however, both successive testing and change in administrators brought about one difference each at the 5 per cent level of confidence. Examiner differences accounted for an additional difference at the 10 per cent level.

o. H: The number of human-figure responses varied at the 10 per cent level once with change in administrators and once with variation in the type of administration.

p. Hd: Human-detail responses were unaffected by retesting or by affective loading of the test situation; however, there was one difference at the 5 per cent level with change in administrators.

q. Ad: Animal-detail responses were also impervious to the emotional tone of the administration and to successive testing, but they twice showed a difference at the 5 per cent level with variation in administrators.

r. CR%: The percentage of responses to the all-color cards remained somewhat stable, showing only a single difference at the 10 per cent level of confidence upon variation of the affective loading of the administration.

s. #W: The number of whole responses varied once at the 5 per cent level with successive testing but remained impervious to change with varied administration.

t. W%: W% remained unchanged with variation in emotional loading of the test situation, but showed two differences at the 1 per cent level with retesting and one difference at the 5 per cent level with variation in administrators.

u. D%: This location scoring symbol varied once at the 5 per cent level with change in examiners but otherwise quite unchanged.

v. d%: The percentage of small usual details remained stable except for one difference at the 10 per cent level with successive testing.

w. S: The number of white-space responses did not differ from one examination to the next but varied between examiners once at the 5 per cent level and differed once at the 10 per cent level with alteration of affective loading of the administration.

x. S + (s): The sum of main and additional white-space responses resisted

change with successive administrations; however, the variation in affective loading produced one difference at the 5 per cent level. Change of administrators proved to be the most significant variable with one difference at the 1 per cent level, one at the 5 per cent level, and a third at the 10 per cent level of confidence.

#### F. REVIEW OF RESULTS RELATED TO ORIGINAL HYPOTHESES

Let us briefly review the specific hypotheses and predictions made at the outset of this experiment and check our results against them.

1. The crucial hypothesis related to the stability of responses to the ink blots regardless of variation in environmental stimuli other than the ink blots; i.e., the person presenting the blots, the implicit attitude of the administrator, or the mere repetition in presentation of the blots. Of the twenty-three Rorschach elements subjected to statistical analysis, only three were relatively unaffected by the extra-test stimuli, by "situational" stimuli. Nineteen of the twenty-three elements entered into differences at least once at the 1 per cent, 5 per cent, or 10 per cent levels of confidence. These nineteen varying factors showed a total of forty-six differences attendant upon some variation in stimuli other than the test stimuli—the ink blots themselves. The hypothesis of stability does not hold.

2. It was predicted at the outset that the number of responses would vary with administrators, regardless of number or affective tone of the administration. The means of responses for the three administrators were, roughly, 24, 27, and 33, the difference between highest and lowest means being significant at the 1 per cent level of confidence. These findings support the prediction.

3. No consistent variation in the number of responses was anticipated from the first to the second to the third administration; and, in fact, there was none. Initial records of over seventy responses were expected to decrease with successive administrations, which they did. However, the prediction that initial records under thirty responses would show successive increase in number of responses was not supported by the data.

4. Stability as a function of individual differences reflected in  $M : \text{sum}C$  ratio proved, as predicted, to isolate approximately 30 per cent of the cases. Also fulfilled was the prediction that the more stable personality records, those of the M-type individuals, would show more situational stability than the records of the C-type persons. Thirty-nine per cent of the subjects retained a stable Experience Type throughout the experiment, approximately three-fourths of whom were M-type personalities, one-fourth C-type.

5. The hypothesis that there would be some variation in the records obtained by the three administrators, regardless of number or type of adminis-



tration, was more than supported by the data. In fact, examiner differences proved the most pervasive influence in the experiment—perhaps the most dramatic finding of the study.

6. The predicted increase in frequency of  $k$ ,  $K$ ,  $c$ , and  $C'$  with the negative administrations was in no particular realized. In fact, both  $F_k$  and  $F_c$  had the lowest mean frequency in the negative testing situation; the other symbols— $K$ ,  $FK$ , and  $C'$ —showed means on the negative administrations which were exceeded by the mean frequencies on the positive administrations. Apparently, as judged by these determinants, our thirty-six subjects found the positively loaded testing situation more challenging or threatening than the negatively loaded. This finding negates the hypothesis under discussion.

7. The hypothesized increase in white-space responses in the negative testing situation was somewhat supported, if one judges negative means against control means. Both main  $S$  and main-plus-additional  $S$  responses showed a mean increase in the negative administrations over the control administrations. However, the negative means of both measures were exceeded by the means on the positive administrations. Apparently the negative administrations did provoke more oppositional behavior than the control administrations; however, the positive administrations elicited even more of this type of response than either the negative or the neutral administrations, a finding which, in effect, negates the hypothesis of peak negativistic behavior on the negative administration.

8. Both crude and refined control, as measured by  $F\%$  and  $FK + F + F_c$  per cent respectively, were expected to increase in the negative administrations. These predictions were supported by the data.

9. The anticipated increase in emotionality on the positive administration, as measured by  $\text{sum}C$ , was realized. There was a steady rise in mean  $\text{sum}C$  from the negative to the neutral to the positive administrations.

10. The increase in movement responses ( $M$  and  $FM$ ) on the positive administrations occurred as predicted. Both scoring categories achieved their highest mean with the positive administration. The lowest mean frequency on  $FM$  occurred with the negative administration; on  $M$ , with the neutral administration.

## *V. Discussion of Results*

In the foregoing report of all of the significant findings for the experimental variables of this study—administrations, administrators, and types of administrations—there was little attempt to make qualitative evaluations or

dynamic interpretations of the findings. Nevertheless, one is justified in attempting—in fact, is obligated to attempt—to relate these data to the dynamics of personality.

#### A. EXPERIMENTAL VARIABLES

An effort will be made to summarize the effects produced upon the experiencing subjects by the different variables herein studied. Too, an attempt is made to understand the differences among the three examiners in terms of the different effects they so consistently produced in the test results. While these procedures have highly speculative components, every effort is made to hold the speculations close to the actual data, to translate into dynamic terms the actual findings of this experiment.

*Influence of repetition of the test.* The mere repetition of the Rorschach was the least important variable in the entire study, accounting for the smallest total number of important *t*-values, only eight in all. The categories showing change with successive administrations are, for the most part, such as can be quite easily accounted for. One is not surprised to find the average time per response consistently decreasing with repetition, this change accounting for two administrative *t*-values. On being successively confronted with the same stimuli, subjects needed no time to adjust to the stimuli; they could employ mere recall as a substitute for perception, apperception, and conceptualization. Furthermore, there is the strong possibility that the subjects, on once learning the total demand or expectancy offered by the test-challenge, simply functioned in line with our cultural speed-value, meeting the total requirements of the situation with as little expenditure of time as could be managed.

We have elsewhere indicated that our subjects frequently showed moderately compulsive tendencies. As a defense mechanism, the compulsive tendencies would be expected to be most operative when the subjects felt most challenged or threatened. One influence of repetition of the test on this dynamic pattern would be reduction of threat or challenge and consequent reduction in display of the compulsive defense. Three of our eight administration *t*-values grow out of reduced use of the whole blot in concept formation on successive tests. This phenomenon would seem to need no further comment.

Repetition of the Rorschach induced a successive and consistent increase in the use of "refined control." Again, one does not have to delve into deep dynamics of personality to understand this finding. Our intelligent subjects merely demonstrated their ability to improve their mastery over a challenging situation with each successive association with the test-task.



The remaining measurable influence of repetition on the test results is concerned with M, human movement responses. On taking the test for the second time, our subjects manifested greater equilibrium, more inner adjustment, a higher level of imaginative thinking. This is in line with the previously noted increase in security, the display of refined control over the situation which attended successive administrations. The third administration of the same test, however, failed to stimulate our subjects to display this sign of the superior, well-adjusted person. There is a measurable decrease in human movement responses on the third administration. One may speculate that the subjects were inhibited on the first administration, were less inhibited and more secure on the next association with the blots, hence free to demonstrate optimal capacity to absorb and creatively utilize stimuli, but were perhaps just bored with the third presentation of the same blots, bored to the point of not bothering to behave optimally.

The suggestion that boredom attended the third testing situation is somewhat supported by the spontaneous comments of many of the subjects during the third session: "These again?" "I've already done this twice." "Same old thing—a butterfly," etc., etc.

*Effect of Examiner A.* Among administrators, A elicited responses which reveal the following functioning of the subjects, as indicated by rank order of means: Fewest number of responses, shortest average time per response, least degree of emotional adjustment or social adjustment (FC), least amount of emotional lability (CF), smallest measure of emotionality (sumC), least expression of sensitivity (Fc), least tendency to use all possible environmental stimuli (C'), least stimulation of critical or perfectionist attitudes toward concepts (Ad and Hd), lowest frequencies in all location areas except W% wherein A's subjects show the highest mean.

The signs on which A's mean frequencies were highest in rank order suggest the following reactions to the effect of this administrator on the subjects: Greatest amount of crude control (F%), most frequent manifestation of unhealthy, uncontrolled direct emotional interaction with the environment (C), strongest tendency toward compulsive behavior (W%).

The pattern which seems to form, as one reviews these summarized effects of Administrator A on the subject's Rorschach records, somewhat resembles what one might expect of a person confronted with a threatening, frustrating situation. If this total effect may be considered a mirror of the administrator's personality, then Examiner A would be described as a cold, forbidding, frustrating, threatening figure, these personality components permeating the test-situation regardless of the deliberate and varied role-playing attempted.

There is no way of checking the validity of this thumb-nail sketch of A's

personality components as reflected in her effect on Rorschach records. The writer, however, did ask two psychologically sophisticated persons, acquainted with all three examiners, to give subjective descriptions of each. Both described A as the coldest, most inflexible, and most solid of the examiners. One added the adjective "masculine" and the phrase "castrating type of female."

*Effect of Examiner B.* Administrator B's results fell in the middle position on rank order of means with the majority of measures analyzed in this study. The relative effect of her personality on the administrations is revealed in part, perhaps, by the very paucity of extreme positions achieved under her administrations. Following are the measurable effects: smallest amount of introspection (FK), lowest indication of maturity or creative imagination (M), and least evidence of buoyancy or free expression of primitive drives (FM).

The subjects, under B's administrations, achieved rank order of 1 on only two significant means of categories: highest total degree of emotionality (sumC) and most refined control ( $FK + F + Fc/R$ ).

The pattern reflected in the mirror of effects suggests that B may possess personality traits that are emotionally exciting to the subjects; at the same time, this excitement apparently is well controlled and does not provoke undue anxiety or tension, neither does it stimulate intellectual activity.

It will be remembered that Administrator B is the youngest of the three examiners. She was subjectively described as the most feminine of the group, the "softest," most nearly the "ideal protecting mother-figure." One person described her as essentially "seductive."

*Effect of Examiner C.* Administrator C stimulated in subjects the least amount of either crude or refined control ( $F\%$  and  $FK + F + Fc/R$ ), the lowest measure of uncontrolled emotionality (C), and the least indication of "whole compulsion" ( $W\%$ ).

The means of the significant measures showing highest rank order under C's administrations are the following: Most responses with shortest average time per response, most introspection and unfixed anxiety (FK and K); best social and emotional adjustment (FC), most evidence of creative imagination (M) and of buoyancy (FM), highest indication of sensitivity (Fc), freest use of all possible stimuli (C'), most evidence of ease in relating to people (H), greatest evidence of critical or perfectionist attitudes toward productions (Ad and Hd), greatest use of all blot location areas except  $W\%$  wherein C's subjects show the lowest mean.

The effect of Examiner C on the results of the Rorschach are more complex than the effects of the other examiners. One may speculate that C provided more of an intellectual than an emotional stimulation to the subjects.



There would seem to be evidence of challenge with attendant anxiety, but there is also evidence of easy rapport, social adjustment, and less need for control devices in the situation with C.

C is the eldest of the three examiners. She was subjectively described as the most flexible of the three, as more feminine than A but less than B. One person described her as exuberant, "bubbling," the other as "very sympathetic."

Before leaving these speculations on the effects of the various examiners, it should be pointed out that Administrator A quite consistently approximates the pattern of the negative administrations with sixteen agreements out of a possible thirty-one. Administrator B most nearly approximates the pattern of the neutral administrations with fourteen agreements out of thirty-one. Administrator C almost duplicates the pattern of the positive administration with twenty-six agreements out of the possible thirty-one.

One may speculate that the basic personalities of the examiners, which permeated all testing situations regardless of the varied affective roles they played, are in fact related: A = negative; B = neutral; C = positive. Or one might infer that Administrators A and B, for example, in deliberately structuring the positive administrations actually created a situation somewhat like that which was consistently and unintentionally created by C in all administrations, that A and C, in attempting to be neutral succeeded only in resembling B's essential personality, and that B and C in structuring negative situations succeeded merely in resembling A's basic pattern. The former interpretation seems more likely. Whichever interpretation appears more reasonable, the facts and figures serve further to underline the importance of examiner differences.

*Effect of positive administrations.* This highly permissive situation was deliberately structured to give subjects the feeling of acceptance; the atmosphere was one of approval from beginning to end. This procedure brought about significant changes in the Rorschach behavior of the subjects, as has been previously noted. Let us summarize the effect of these positive administrations:

More responses were elicited, suggesting greater productivity in the more permissive situation. These responses tended to incorporate all possible stimuli presented. There was a greater display of intellectual activity, of creative imagination, plus evidence of increased thinking of a popular or communal type. Stereotypy of thinking was markedly low during the positive administrations. There was evidence of increased resistance against feelings of personal inadequacy plus a rise in tension-tolerance and in self-questioning, or introspection. Finally, there was marked evidence of a greater ease in relating to other human beings.

*Effect of negative administrations.* The negative test-situations were non-permissive, rejecting; an atmosphere of disapproval prevailed throughout. The effect of this environmental tone on the Rorschach records is reflected in a definite increase in stereotypy of thinking and a decrease in thinking that is in harmony with community thinking. There is a low incidence of intellectual activity of the imaginative or creative sort. Too, there is evidence of withdrawal from emotional stimuli, as measured by a decrease in the number of responses to the all-color cards. There is a marked rise in self-questioning, or introspection, plus evidence of resistance against feelings of personal inadequacy.

*Effect of neutral administrations.* The chief value in the experimental design of the neutral set of Rorschachs is that of a control factor. Since these tests were administered in the standard manner, without affectively loaded pre-tests, they serve somewhat as a measure of the reactions to the Rorschach which one might expect to find under ordinary clinical situations. The results on the neutral Rorschachs, however, cannot be construed to have the same meaning as is ordinarily attributed to the results of control groups in psychological experimentation. A third of our subjects had had one affectively loaded Rorschach before experiencing the neutral situation; two thirds had had two affectively loaded experiences with the test before receiving the neutral administration. It would be folly to assume that all of these neutral administrations provided identical test-situations for our subjects, that there was no carry-over from previous experiences with the administrative tone in which the blots were presented. Nevertheless, our neutral administrations do offer a modicum of evidence as to how subjects react to the test when standard testing conditions prevail.

The neutral situation elicited fewer responses than the positive, and these responses showed higher stereotypy of content than occurred on the positive administrations. There was less interest in utilizing all environmental stimuli, diminished ease in interpersonal relationship, and diminution of intellectual creativity on the neutral—as compared with the positive—administrations. Lower tension-tolerance and less resistance to feelings of inadequacy were also apparent in the neutral, over the positive, administrations. A lower resistance reaction was evident on the neutral administrations when compared with the negative situations. Less introspection was apparent on the neutral than on the positive administrations, and there was greater reaction to the emotional stimulation of all-color cards.

#### B. IMPLICATIONS FOR FURTHER INVESTIGATION

This study does not supply the answer to a question which grows out of the findings: What kind of personality was possessed by each of the three



administrators? Since administrator differences appeared as the greatest contributing factor to the variation in the Rorschach records, one would like to know what specific and measurable factors in an examiner's personality will bring about what specific and measurable factors in a subject's Rorschach protocol. One wonders if a Rorschach psychogram of the three examiners would resemble the composite mean profiles of the subjects in every way or in any way. These problems should be investigated.

The present study was designed in such a way that a subject could withdraw without suffering too great a penalty. Consequently, several of the individuals who reacted intensely to the negative administration withdrew from the project without finishing the three tests. Their records, of course, could not be included in this study. This entire experiment should be repeated with a group of subjects over whom the examiner can exercise greater control. The fact that differences were revealed as a function of the negative administration makes one wonder how much more significant these differences might have been—and what additional differences might have appeared—if the records of the intensely affected subjects could have been incorporated in the data.

The experimental design of the present study required that each subject not only experience three different affective loadings of the test situation but also experience three different examiners. Two separate investigations should be made as a supplement to the present study. It would be well to know whether the same kind and amount of fluctuation of scoring symbols and M: sumC balance would occur if a single examiner tested a group of subjects under varied affectively loaded situations or if several examiners successively tested a group of subjects under the same affectively loaded circumstances. This latter study should, in fact, be broken into at least two studies, one investigating the results of a series of positively loaded situations, and the other exploring the results of a series of negatively loaded administrations.

In order to minimize the personal equation and thereby to study in purer form the influence of negative and positive rapport factors on Rorschach performances, one could repeat this experiment with the following variations: neutral, positive, and negative administrations could be administered in group form, according to the design herein used, with directions or instructions given by a single examiner through the medium of phonographic recordings. This examiner would, of course, have to cut three records, each affectively differently loaded. While such a procedure could not effectively incorporate the particular card-sorting pre-tests used in this study as tone-setters for the administration, other equally simple devices could be

created. In fact, the findings of the present study suggest that no such elaborate preamble is necessary to instill feelings of acceptance or rejection. A cold reception proved adequate to disturb our subjects to the point of misunderstandings and errors.

The significant findings for the group of male college sophomores justify a repetition of this entire experiment with other age groups and with female subjects. While our subjects were randomly selected, they were not selected from the population as a whole but from a rather small stratum of a very small and highly selected population—college sophomores. Our group was shown to deviate from population norms in several instances; e.g., in compulsive tendencies and pedantic approach. One may only speculate on the extent to which other findings in this experiment are a function of the select group from which the subjects were drawn. Would a similar study, with subjects representing a cross section of the population as a whole, show more or less variation, show the same or different kinds of instability?

All of these questions, and many more, must eventually be answered in the laboratory if we are to refine our use of the most popular and pragmatically valuable of the projective techniques: The Rorschach Ink Blot Test. This study is a small step in a direction which must, and certainly will, be taken by other students who realize that the practical success of the Rorschach must be supported by laboratory verification. This area is rich in experimental problems sufficient to excite at least a generation of students.

## *VI. Summary and Conclusions*

This investigation has attempted to explore the stability or variability of the numerous response-determinants of subjects' concept on the Rorschach Ink Blot Test, to discover the extent to which mere repetition of the test, a change of administrators, or an alteration of the affective tone of the testing situation would bring about alterations in number and content of responses, in determinants, and in location areas.

Related studies are extremely few and their contradictory results are explainable as a function of their decidedly different experimental designs. A single study, that of the Army Air Forces, is closely related to the present experiment in its analysis of variation in number of responses as a function of examiner-differences.

What other response categories are subject to fluctuation as a function of examiner-differences? If examiners studiously attempt to structure three different affectively loaded testing situations, will there be consistent differences, or any differences, as a function of these deliberately varied rap-



port situations? Will mere repetition of the Rorschach significantly alter any of the scoring categories, regardless of administrator or of affective loading of the administration? If so, which scoring categories are sensitive to such variation?

Thirty-six male college sophomores between the ages of nineteen and twenty-seven participated in a series of three Rorschach tests, administered without testing of the limits. Each test was administered by a different female, and each testing situation was affectively differently loaded. One test was given in the standard manner without pre-tests and without any attempt to give the subject a marked feeling of acceptance or rejection. Another test was administered following two pre-tests during which, by manner and tone of voice, the subject was made to feel rejected and a failure. Still another test was administered after a similar pre-test situation during which the subject was made to feel accepted and successful. Each examiner gave twelve tests first, second and third; twelve tests positively loaded, negatively loaded, and neutral without affective loading. Each subject's different test was administered by a different examiner. An equal number (twelve) of negative, positive, and neutral tests was given on the first, second, and third administrations and by the three different examiners. This plan netted thirty-six Rorschach protocols on each of the three administrations, thirty-six by each of the different examiners, and thirty-six under each of the three different affective loadings of the testing situation. The 108 Rorschach protocols were scored and separately analyzed according to the number of the administration, the administrator, and the type of administration.

On the basis of the analyses of the 108 Rorschach records, the following findings are reported: Subjects who were coldly received tended to misunderstand, and make errors in following directions on simple card-sorting tests. Subjects warmly greeted did not misunderstand, request repetition of, or make errors in following simple directions.

The Experience Balance, or  $M : \text{sum}C$  ratio, was unstable, only 30% of the subjects maintaining a constant balance throughout the variations in administration. The remainder shifted emphasis from one side of the ratio to the other at least once.

Ten, thirty-one, and forty-eight  $t$ -ratios were found at the 1 per cent, 5 per cent, and 10 per cent levels of confidence respectively. Of these forty-eight measurable differences, eight were a function of repetition of the test, thirteen a function of the variation in affective loading of the test situation, and twenty-seven a function of examiner differences.

Of the twenty-three Rorschach response-categories subjected to statistical analysis, three were relatively unaffected by the extra-test, or "situational" stimuli:  $Dd + S\%$ ,  $Fk$ , and  $A$ . The other twenty categories showed fre-

quency differences at least once at the 1 per cent, 5 per cent, or 10 per cent levels of confidence under at least one of the experimentally varied situations.

In conclusion, performance on the Rorschach, as measured by the frequency of responses in the Rorschach categories, varies significantly with repetition of the test, with variation in the negative or positive rapport conditions of the administration, and with examiner differences. In the present study, the largest and most frequent variations in Rorschach performance were associated with examiner differences.

## APPENDIX

### *Explanation of Scoring Symbols Used in this Study\**

LOCATION		POPULARITY	
W	Whole blot	P	Popular responses
W, S	Whole blot and white space used (tabulated as main W and additional S)		
D	Large usual detail		DETERMINANTS
D, S	White space used in addition to D (tabulated as main D and additional S)	R	Response
d	Small usual detail	M	Figures in human-like action
Dd	Unusual detail (or unusual combinations of usual areas)	FM	Animals in animal-like action
S	White space	m	Abstract or inanimate movement
		k	Shading as three dimensional expanse projected on a two dimensional plane
		K	Shading as diffusion
		FK	Shading as three-dimensional expanse in vista or perspective
		F	Form only, not enlivened
		Fc	Shading as surface appearance or texture, differentiated
		C'	Achromatic surface color
		FC	Definite form with bright color
		CF	Bright color with indefinite form
		C	Color only
			$\text{sumC} = \frac{\text{FC} + 2\text{CF} + 3\text{C}}{2}$
CONTENT			
H	Human figures		
Hd	Parts of human figures, not anatomical		
A	Animal figures		
Ad	Parts of living animals		

\* After Klopfer, Bruno and Helen H. Davidson, *The Rorschach Method of Personality Diagnosis*. Individual Record Blank (New York: World Book Company, 1942).

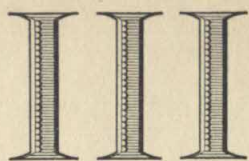


## REFERENCES

1. BECK, S. J. *Rorschach's test: I Basic processes*. New York: Grune & Stratton, 1944.
2. BECK, S. J. *Rorschach's test: II. A variety of personality pictures*. New York: Grune & Stratton, 1945.
3. BELLAK, L. The concept of projection. *Psychiatry*, 1944, 7, 365-366.
4. BERGMAN, M. S., GRAHAM, H., & LEAVITT, H. C. Rorschach exploration of consecutive hypnotic chronological age level regressions. *Psychosom. Med.*, 1947, 9, 20-28.
5. BOCHNER, RUTH, & HALPERN, FLORENCE. *The clinical application of the Rorschach test*. New York: Grune & Stratton, 1945.
6. BRUSSEL, J. A., & HITCH, K. S. The Rorschach method and its uses in military psychiatry. *Psychiat. Quart.*, 1942, 16, 3-27.
7. BUHLER, CHARLOTTE, BUHLER, K., & LEFEVER, D. W. *Development of the basic Rorschach score*. Los Angeles: mimeographed ed., 1948.
8. BUHLER, CHARLOTTE, & LEFEVER, D. W. A Rorschach study on the psychological characteristics of alcoholics. *Quart. J. Stud. Alcohol*, 1947, 8, 197-261.
9. COFER, C. N. Psychological test performance under hyoscine: A case of post-infectious encephalopathy. *J. Gen. Psychol.*, 1947, 36, 221-228.
10. DOUGLAS, ANNA GERTRUDE. A tachistoscopic study of the order of emergence in the process of perception. *Psychol. Monogr.*, 1947, 61, No. 6.
11. FOSBERG, I. A. How do subjects attempt fake results on the Rorschach test? *Rorschach Res. Exch.*, 1943, 7, 119-121.
12. FOSBERG, I. A. Rorschach reactions under varied instructions. *Rorschach Res. Exch.*, 1938, 3, 12-31.
13. HALL, CALVIN S. Diagnosing personality by the analysis of dreams. *J. Abnorm. Soc. Psychol.*, 1947, 42, 68-79.
14. HEIDER, F. Social perception and phenomenal causality. *Psychol. Rev.*, 1944, 51, 358-374.
15. HERTZMAN, M., & PEARCE, JANE. The personal meaning of the human figure in the Rorschach. *Psychiatry*, 1947, 10, 413-422.
16. HSÜ, E. H. The Rorschach responses and factor analysis. *J. Gen. Psychol.*, 1947, 37, 129-138.
17. HUNT, W. A. The future of diagnostic testing in clinical psychology. *J. Clin. Psychol.*, 1946, 2, 311-317.
18. JACOB, Z. Some suggestions on the use of content symbolism. *Rorschach Res. Exch.*, 1944, 8, 40-41.
19. KADINSKY, D. Human whole and detail responses in the Rorschach test. *Rorschach Res. Exch.*, 1946, 10, 140-144.
20. KAMMAN, G. R. The Rorschach as a therapeutic agent. *Amer. J. Orthopsychiat.*, 1944, 14, 21-27.
21. KLOPPER, B., & KELLEY, D. M. *The Rorschach technique*. New York: World Book Co., 1946.

22. LEVINE, K. N., GRASSI, J. R., & GERSON, M. J. Hypnotically induced mood changes in the verbal and graphic Rorschach: A case study. *Rorschach Res. Exch.*, 1943, 7, 130-144.
23. LEVINE, K. N., GRASSI, J. R., & GERSON, M. J. Hypnotically induced mood changes in the verbal and graphic Rorschach. Part II: The response records. *Rorschach Res. Exch.*, 1944, 8, 104-124.
24. LINDNER, R. M. Content analysis in Rorschach work. *Rorschach Res. Exch.*, 1946, 10, 121-129.
25. LINDNER, R. M. Some significant Rorschach responses, *J. Crim. Psychopathol.*, 1944, 5, 775-778.
26. LINDQUIST, E. F. *Statistical analysis in educational research*. Boston: Houghton Mifflin, 1940.
27. LUCHINS, A. S. Situational and attitudinal influences on Rorschach responses. *Amer. J. Psychiat.*, 1947, 103, 780-784.
28. PIOTROWSKI, Z. A comparative table of the main Rorschach symbols. *Psychiat. Quart.*, 1942, 16, 28-37.
29. RORSCHACH, H. *Psychodiagnostics. A diagnostic test based on perception*. Berne, Switzerland: Hans Huber, 1942. English ed. by Paul Lemkau and Bernard Kronenberg.
30. SCHACHTEL, E. G. Subjective definitions of the Rorschach test situations and their effect on test performance. *Psychiatry*, 1945, 8, 419-448.
31. SCHACHTEL, E. G. Review, *The Rorschach technique*, by Bruno Klopfer. *Psychiatry*, 1942, 5, 604-606.
32. U.S. ARMY AIR FORCES AVIATION PSYCHOLOGY PROGRAM RESEARCH REPORT, J. P. GUILFORD, editor. *Printed classification tests*. Report No. 5. Washington, D.C.: Government Printing Office, 1947. Chapter 24, Clinical type procedures.
33. WILKINS, W. L. & ADAMS, A. J. The use of the Rorschach test under hypnosis and under sodium amytal in military psychiatry. *J. Gen. Psychol.*, 1947, 36, 131-138.





## *Scoring*

J. R. Wittenborn

STATISTICAL TESTS OF  
CERTAIN RORSCHACH  
ASSUMPTIONS , *Analyses of*  
*Discrete Responses*

THE CUSTOMARY procedures employed in the evaluation of a subject's responses to the Rorschach ink blots involve a variety of assumptions. Some of the assumptions are explicitly stated in the Rorschach literature and others are merely implied by common practices. Although the validity of the Rorschach technique for personality appraisal is indicated by clinical findings and certain types of studies, the assumptions employed in the Rorschach psychodiagnosis have not been subjected to direct validation. There are considerations which make particularly important the validation of the assumptions per se.

1. Since the assumptions have not been tested, the possibility exists that some of them are invalid. If some of the assumptions cannot be validated,

Reprinted from *J. Consult. Psychol.*, 1949, 13, 257-267, by permission of the American Psychological Association and the author.



aspects of scoring or interpretation based on the questionable assumptions should be employed with reservation or perhaps eliminated from standard procedures.

2. Many of the assumptions employed in Rorschach procedures cannot be found among the established facts and theories of American academic psychology. If the Rorschach assumptions are valid, their relevance to the psychology of personality organization and to the psychology of perception could be of extraordinary significance.

*Analysis A. An Examination of the  
Functional Similarity of Responses  
Which Have Common Determinants*

In scoring a Rorschach protocol, it is customary to classify each response with respect to several classes of factors, such as content, portion of ink blots interpreted, and ink-blot characteristics determining the response. There are standard categories within each of the classes of factors. When a response is scored the content category in which it falls, as well as its location category and its determinant category, is designated by the scorer. As a second step in scoring, the total number of responses falling in each category is ascertained. Regardless of the precise manner in which the categories are used, the device of categorizing the responses and then finding the total number of responses in each category carries two implications:

- I. That all of the responses falling in a given category are similar in some behavioral respect.
- II. That the psychological significance of responses falling in a given category is different in some respect from responses placed in other categories.

These assumptions are implicit for example in the distinction that is made between color and human movement responses. The importance of the distinction between human movement and color responses was first emphasized by Rorschach, and the observance of this distinction has remained a basic feature of Rorschach psychodiagnosis. Because of the basic significance ascribed to these two types of responses, it was decided to employ this distinction in testing the foregoing assumptions of intracategory similarity and intercategory difference.

It would be erroneous to suppose that any absolute or relative quantity of color or movement responses is unalterably and invariably associated with

a particular total personality appraisal. Movement and color have an invariable abstract meaning for the personality, but for total personality appraisals this is added to, detracted from, or modified by other characteristics of the response and other aspects of the protocol. This is analogous to clinical practice wherein the psychologist's evaluation may be based on the Rorschach indications, other test data, and data from the case history; the relative contribution ascribed to the Rorschach in the total personality evaluation is determined to some degree by the other findings, but this does not render a Rorschach finding meaningless. By the same token the qualifying role of such factors as form level and content do not make the concept of a movement response meaningless nor make inconsequential an assumption that movement responses include a constant, functionally similar element.

#### THE PLAN OF EXPERIMENT

If Assumption I is a true proposition, it follows that several measurements of the tendency to perceive human movement would be positively related. If several measures of a given response category were not positively interrelated, it would be difficult to justify an assumption that differences in number of responses of that category expressed differences in a personality attribute. It also follows from Assumption II that several measures of a given response category would be more highly interrelated with each other than they would be related with measures of some other category.

In order to employ these deductions in an experimental test of the assumptions, it is necessary to devise several measures of the tendency to give responses falling in the respective categories. For example, it is possible to regard any response falling within a given category as evidence of a tendency to give responses in that category, and it is equally possible to regard failure to give a particular response falling in a given category as evidence of a tendency not to give responses falling in the category. Employing this approach, it is possible to secure a large number of plus or minus scores for any response category. If making or not making a particular response which belongs to a given category is to be a measure for a tendency to make respectively many or few responses which fall in that category, it is important that the particular response be well specified. This poses an important methodological problem.

One of the outstanding characteristics of Rorschach responses is their highly individual quality, their extreme variability from person to person. This makes conventional statistical treatment of Rorschach data awkward because of the difficulty in securing a sufficiently large number of roughly identical responses. For example, out of a group of one hundred normal



subjects many, if not all, would report human movement responses, but relatively few would have more than two or at the most three identical responses in common.

In the present study an attempt has been made to reduce this difficulty by use of the Harrower-Erickson check list (6). This check list comprises a sheet of paper upon which are printed three groups of ten possible responses for each of the ten cards. There are three hundred possible responses in all. The technique of administration provides that the ten cards be projected on a screen in the usual serial order. The check list is designed so that the subject may check or underline responses which indicate what he has seen on the card. Although this technique of administration and manner of responding to the cards is obviously different from the standard technique for administering the Rorschach test, this device provides a uniform, common response list for all subjects and thus makes possible a tabular treatment of responses to ink blots. The limitation of this procedure as a method for personality appraisal (3, 5, 7) is not necessarily disadvantageous to the present experiment which is an examination of the consequences of certain restricted assumptions and is not otherwise concerned with the problem of the validity of the Rorschach method in general or the validity of other Rorschach assumptions.

The following experimental hypotheses were tested by a statistical analysis of data provided by the Harrower-Erickson check list:

A. The tendency for controlled color responses (FC and CF) to be associated with other controlled color responses (FC and CF) is greater than the tendency for such color responses to be associated with human movement responses.

B. The tendency for human movement (M) responses to be associated with other human movement (M) responses is greater than the tendency for such responses to be associated with controlled color responses (FC and CF).

## RESULTS

The subjects for the investigation were 247 Yale University freshmen, all the members of a small entering class. The check list was administered as a part of a guidance battery. The responses selected for analysis had to meet several requirements. The responses had to be sufficiently common to afford a basis for statistical tests. The permissible range of response frequency was arbitrarily set from 28 to 220. The exact range was influenced by efforts to meet certain other requirements, e.g., an effort was made to include responses from all cards; different responses to the same portion of the card

which could therefore be construed as mutually exclusive were (with two noted exceptions) avoided. The responses finally selected are listed in Table I.

The color responses comprise two groups which differ in the degree to which they are influenced by the form characteristics of the card. Responses numbers 8, 9 and 10 are strongly influenced by form and would be scored FC, whereas responses 11, 12 and 13 are less influenced by form and would be scored CF by most workers. The color responses are not strictly comparable with respect to the location of the response; for example, responses 8 and 10 clearly involve large details of the cards, whereas with the exception of response 9, which is probably based on the whole card, the others are less definite in their exact location. For the most part the human movement responses are based on most if not all of the total area of the respective ink blots. Responses 1, 6 and 7 are based on colored cards and they may involve color to some degree. This possibility is greatest for response 1; clowns (or equivalent movement responses for card II) are commonly given a specific identification which employs the red color. It is possible to control these complicating and extraneous factors by restricting comparisons to certain of the selected responses. For example, human movement responses 3, 4 and 7 comprise practically all of the ink blot and cannot involve color because the blots are achromatic. Similarly, responses 11, 12 and 13 are homogeneous in that they tend to involve most of the card and would usually be scored CF.

In order to test the experimental hypotheses, each response was correlated with every other one. Two different statistical devices were employed to examine the relationships. The  $\chi^2$  test of independence was first used. The size of  $\chi^2$  is not determined by the degree of relationship, but since the frequency with which  $\chi^2$  of any given size may occur by chance is known,  $\chi^2$  may be used as a test for the significance of the correlation, i.e., the lack of independence between two responses may be evaluated by means of the  $\chi^2$  test. The result of this analysis is shown in Table I.

In order to get some appreciation of the strength of the significant relationships revealed by the  $\chi^2$  test, tetrachoric correlations between the responses were computed. The correlations were determined by means of a chart (4). Since the tetrachoric correlation is an appropriate statistic only when the dichotomized variables are normally distributed, its use for the present purpose may be questioned; whether the strength of a tendency to give a Rorschach response is normally distributed or not is a speculative question. At any rate the indications provided by the tetrachoric correlations are in agreement with those provided by the  $\chi^2$  test. The tetrachoric correlations between all possible pairs of responses are given in Table II.





TABLE II  
Tetrachoric Correlations Between Rorschach Responses

No.	Response	Card No.	Resp. Group	1	2	3	4	5	6	7	8	9	10	11	12
1	Two Clowns	II	a												
2	Two Men Pulling Something Apart	III	a	.11											
3	Man Seen From Below	IV	c	.20	.20										
4	A Fan Dancer	V	c	.21	.16	.42									
5	Two Women Talking	VII	a	.32	-.01	.16	.17								
6	Two Witches	IX	a	.20	.20	.22	.23	.26							
7	Two People	X	a	.08	-.10	.40	.27	.30	.33						
8	A Red Bow Tie	III	a	.00	-.14	.30	.37	.15	.43	.23					
9	An Emblem	VIII	a	.08	-.07	.13	.14	.02	.15	.30	.15				
10	Colored Map of California	X	a	.00	-.05	.14	.24	.03	.20	.10	.20	.22			
11	Red and Black Ink	II	a	-.38	.01	-.10	-.10	-.50	.07	.06	-.01	.02	.13		
12	Fire and Ice	VIII	b	.33	.05	.34	.16	.20	.40	.33	.31	.07	.31	.09	
13	Forest Fire	IX	b	.20	-.05	-.07	.44	.20	.07	.22	-.10	.21	.02	-.15	.20



## DISCUSSION

Two conclusions seem equally appropriate from an examination of either Table I or Table II:

1. Among the intercorrelations there is no obvious pattern which conforms with a pattern predictable from either Hypothesis I or II.
2. The number of significant correlations exceeds the number that would be expected by pure chance.

Although it would be unwarranted on the basis of the present data to make any conclusive statements concerning the general status of Assumptions I and II, the hypotheses generated by them for this experiment are not clearly verified. If the Rorschach were analogous to the typical mental test wherein the score is based on many responses of a given kind and evaluation of the individual is based on differences in many score units, a high degree of interitem consistency would not be necessary. But such is not the case, and in view of the considerable psychological significance ascribed to small variations in number of human movement or color responses, highly consistent interresponse evidence for the hypotheses may be expected. The scant evidence for the assumptions is a challenge to their status and strongly indicates the need for further investigation.

The fact that among the responses there are numerous relatively large and significant interrelationships which were not predicted by the hypotheses indicates need for a broader examination of the patterning of Rorschach responses. It is possible that the interrelationships among Rorschach responses have a well defined and meaningful pattern which if discovered would contribute to our understanding and use of the Rorschach technique.

Although from a casual examination of Tables I and II it appears that the evidence which the present data provide for the assumptions is unequal to the important manner in which they are employed in practice, any evidence of a trend toward correspondence between the requirements of the hypotheses and the nature of the data should not be overlooked.

Accordingly, Table III was prepared to summarize the findings and from this summary several statements may be offered:

1. There is no consistent functional similarity among the human movement responses or among the color responses, i.e., most of the movement responses are not significantly interrelated and similarly most of the color responses are not significantly interrelated.
2. The pattern among the most significant relationships is in a direction predictable from the hypotheses.

TABLE III

## Distribution of Chi-Squares and Tetrachoric Correlation Coefficients

$\chi^2$	Per Cent of the 42 Possible Comparisons of Human Movement with Human Movement	Per Cent of the 42 Possible Comparisons of Human Movement with Color*	Per Cent of the 30 Possible Comparisons of Color with Color
Less than 2.71	42	69	67
2.71 and above (10% level)	58	21	33
3.84 and above (5% level)	38	19	27
6.635 and above (1% level)	19	15	13
Tetrachoric $r$			
Less than .20	33	62	60
.20 or above	67	38	40
.30 or above	24	21	13
.40 or above	10	05	00

\* Two of the significant  $\chi^2$  are for negative relationships (see Tables I and II).

3. The hypotheses are not clearly verified and the assumptions may be false.

As a result of the analysis it is concluded that in these data the tendencies required by hypotheses A and B are negligible. A variety of considerations may be adduced to account for negligible evidence for experimental hypotheses. Among them the following may be included:

1. The assumptions may be at fault.

a) The assumptions may be false. (The present findings merely challenge the assumptions.)

b) The assumptions may describe a very slight trend. (The present evidence does suggest that the trend is much less marked than might be inferred from Rorschach practices.)

c) The statement of the assumptions may be faulty. (One of the purposes of research is to arrive at correctly stated assumptions, and the refutation of hypotheses is the basis for reformulation of assumptions.)

2. The experimental hypotheses may not be the logical consequences of the assumptions.

a) The hypotheses may involve assumptions other than those under examination. (The present experiment does involve assumptions regarding the meaning and scalability of a response.)



- b) The hypotheses may be incorrectly deduced, e.g., so as to be in conflict with other possible deductions.
3. The experimental interpretation may be faulty.
- a) The data may be partially or wholly irrelevant to the hypotheses. (The present data may be inappropriate to a general examination of Rorschach assumptions, but they are considered appropriate to the present assumptions.)
- b) The data may be influenced by constant, uncontrolled effects (either the group situation or the list of possible responses may have an untoward effect on the individual's responses).
- c) Relative to the magnitude of the trend to be examined, errors of measurement and other variable extraneous factors may be large and thus obscure the findings. (This possibility is not eliminated in the present investigation, e.g., other aspects of the personality not expressed through movement or color may vary from response to response and from individual to individual.)

*Analysis B. An Examination of the  
Functional Similarity of a Large  
Number of Check List Responses*

The major conclusions of Analysis A may be stated as follows:

1. Among the intercorrelations there is no obvious pattern which conforms with a pattern predictable from either Hypothesis I or Hypothesis II.
2. The number of significant correlations exceeds the number that would be expected by pure chance.

These conclusions, particularly the latter one, provide the point of departure for a second analysis. Since there is a greater number of significant intercorrelations among the Rorschach responses than chance alone would provide, it is possible that the pattern of the intercorrelations could be anticipated to some degree from a consideration of certain beliefs and practices concerning the Rorschach. As indicated in the discussion of the results of Analysis A, the consistent behavioral significance among human movement or among color responses may be relatively small compared with the variations among the responses due to other inconsistent factors, e.g., a group of responses selected because they all involve color may differ greatly in other respects. For example, some of the responses will involve the whole card whereas others will be based on a small detail, and the contents of some of the responses may be man-made objects while others may be plants, etc. It

may be reasoned, therefore, that the intercorrelations among responses which have more than one scoring category in common will be higher than the intercorrelations among responses which have but one or no common scoring category. These considerations lead to a third hypothesis for testing the provisions of Assumptions I and II.

Hypothesis C: The degree to which Rorschach responses will be intercorrelated is a function of the number of scoring categories the responses have in common.

Since it is obviously infeasible to study the interrelationships among all of the 300 responses possible in the Harrower-Erickson check list, some selection was necessary. In selecting the responses for Analysis II several factors were considered:

1. The responses must be sufficiently common to make statistical analysis feasible.
2. The manner in which they are scored must be unambiguous.
3. They must be pertinent to common Rorschach practices and assumptions.
4. Whenever possible, the responses to be studied should be selected from the first group of 10 possible responses for each card or at least the response of a given type selected for a card was the first one listed for each card.
5. Within these restrictions the responses selected should represent a variety, and not be limited to one or two types.

As a result of these considerations, twenty-seven check list responses were selected from the records of a small Yale freshman class of 247 and intercorrelated by the tetrachoric method.

As a second step in the analysis, a location, determinant, and content scoring was made for each response. The scoring for each response was then compared with the scoring for every other response and for each response four lists were prepared:<sup>1</sup>

- a) The responses which agreed in three respects (had location, determinant, and content in common).
- b) The responses which agreed in two respects.
- c) The responses which agreed in one respect.
- d) The responses which agreed in no respect (had no common scoring category).

On the basis of these four lists, four distributions of correlation coefficients were made:

- a) The three-degree-of-correspondence distribution comprised all the correlations between each response and every other response which had identical scoring categories (Table IV).
- b) The two-degree-of-correspondence distribution comprised all the correlations

1. The correlations between responses based on a common card were not included in the analysis.



between each response and every other response wherein two scoring categories were shared (Table IV).

c) The one-degree-of-correspondence distribution comprised all the correlations between each response and every other response having but one common scoring category (Table IV).

d) The zero-degree-of-correspondence distribution comprised all of the correlations between each response and every other response having no common scoring category (Table IV).

TABLE IV  
Frequency Distributions of Correlation Coefficients Between Responses  
with Varying Numbers of Aspects in Common

Frequency Distributions of Correlation Coefficients Between Responses With					
Interval of $r$	Three Aspects Same	Two Aspects Same	One Aspect Same	No Aspect Same	Total Distr.
+4 - +5	2	4	14	4	24
+3 - +4	12	6	18	18	54
+2 - +3	20	18	70	22	130
+1 - +2	28	10	74	50	162
.0 - +1	14	12	92	36	154
-.1 - .0	8	4	42	18	72
-.2 - -.1	—	2	28	10	40
-.3 - -.2	—	—	—	—	—
-.4 - -.3	—	—	—	—	—
-.5 - -.4	—	—	2	—	2
-.6 - -.5	—	—	—	—	—
-.7 - -.6	—	—	—	—	—
-.8 - -.7	—	—	—	—	—
-.9 - -.8	—	—	2	—	2
	N = 84	N = 56	N = 342	N = 158	N = 640
	M = +.176	M = +.184	M = +.108	M = +.130	M = +.129
	S = .123	S = .147	S = .168	S = .143	S = .176

Hypothesis C is tested by means of the  $\chi^2$  test shown in Table V. This table shows that there is a significant tendency for the number of response categories common to pairs of Rorschach responses to be positively related with the size of the correlations between the paired responses. The trend does not appear to be a continuous one, however, and as a means of examining it more minutely, each degree of correspondence was compared with every other one. Responses which have two or three common scoring categories have probable higher intercorrelations than responses which have one or less common scoring categories. There is no significant difference between the size of the correlations between responses which have two common categories and the responses which have three common categories: nor

is the difference between the size of the correlations between responses having one common category and the size of the correlations between responses having no common categories significant. These findings are taken as evidence for Hypothesis C.

TABLE V  
Chi-square Between Values of Correlation Coefficients and Number  
of Response-Aspects in Common

Value of r	<i>Number of Respects in which Responses are Same</i>				Total
	3	2	1	0	
	Frequency	Frequency	Frequency	Frequency	
.20 plus	34 (28)	28 (19)	102 (114)	44 (52)	208
.0 - .2	42 (41)	22 (27)	166 (167)	86 (77)	316
Less than .0	8 (15)	6 (10)	74 (61)	28 (28)	116
Total	84	56	342	158	640
	n = 6	$\chi^2 = 18.932$		P = .01	

The relatively small differences between the size of the correlations in each of the four distributions (Table IV) and the variability of each of the distributions is meager support for the practice of categorizing the small number of responses provided by the usual Rorschach protocol. The large number of zero or negative correlations between responses which have two or even three scoring categories suggests that the trend provided by Assumption I is indeed weak. The number of correlations above .30 between responses which have one or even no common category suggests that the distinction between responses having different scoring categories (Assumption II) could be of little or no practical value in appraising individuals on the basis of these data.

## DISCUSSION

As a result of the Analysis B two conclusions are permissible:

1. There is a statistically significant tendency for Rorschach responses which have two or three scoring categories in common to be more highly intercorrelated than responses which have but one or no common scoring categories. Thus Hypothesis C is verified.

2. The tendency for the correlations between responses to vary directly with the number of common scoring categories is no negligible as to be associated with a large number of gross exceptions. The number and size of the exceptional correlations is such as to challenge the propriety of the prac-



tice of scoring the Rorschach check list responses on the basis of location, determinant, and content.

The present study has been concerned with the functional similarity of responses which are scored in a highly similar if not identical manner. The relevance of this analysis is based on the presumption that responses which receive an identical designation on the basis of a complex categorization should have a strong, consistent, demonstrable, functional similarity if the categorization has implications for making distinctions between people. It is similarly argued that responses which are placed in dissimilar scoring categories should have a relatively low degree of functional similarity. The result of the present analysis is not in conformance with these expectations, and the value of the practice of scoring Rorschach responses on the basis of location, determinant, and content is hereby questioned. It is important to note that the data employed in the present study were secured by the use of a check list was necessary to a study such as the present one, and it is to be emphasized that the inability of the examiner to make qualitative judgments of the examinee's behavior, to make an inquiry, test limits, and conduct a searching interview with the client regarding his background and present difficulties (such interviews appear to be the rule, rather than the exception with Rorschach examiners) is beside the point of the present study. There are, however, two possible respects in which the check list technique would elicit atypical data:

1. Actually seeing the possible responses may in some cases suggest a response which would not be made in the standard situation.
2. The presence of others in the testing situation may result in responses which might otherwise not appear.

These possibilities impose a limitation in the degree to which the present results may be generalized to the standard individual Rorschach procedure.

### *Summary of Analyses A and B*

*Analyses A and B* comprise an examination of the responses of 247 students to a Rorschach check list. The analyses were designed to determine empirically the status of certain hypotheses generated by two assumptions which are relevant to practices in Rorschach scoring and interpretation. The assumptions from which the hypotheses were deduced are:

*Assumption I*, All the responses falling in a given Rorschach scoring category are similar in some behavioral respects.

*Assumption II*, The psychological significance of responses falling in a given scoring category is different from that of responses in other categories.

*Analysis A* was designed to examine the status of Hypotheses A and B. The analysis of the data offers negligible support for the validity of these hypotheses:

*Hypothesis A*, The tendency for controlled color responses (FC and CF) to be associated with other controlled color responses (FC and CF) is greater than the tendency for such color responses to be associated with human movement responses.

*Hypothesis B*, The tendency for human movement (M) responses to be associated with other human movement (M) responses is greater than the tendency for such responses to be associated with controlled color responses (FC and CF).

*Analysis B* provided a test for Hypothesis C:

*Hypothesis C*, The degree to which Rorschach responses will be inter-correlated is a function of the number of scoring categories the responses have in common.

The evidence for Hypothesis C met a criterion for a high order of statistical significance. Nevertheless, the trend was slight and marked by numerous important exceptions (see Table IV). Although the data support Hypothesis C they may be interpreted as comprising a challenge to the clinical value of current Rorschach scoring procedures rather than as an indication that the procedures are economical and highly efficacious in attempts at appraising the individual.

The data employed in the present study are check-list data. Had they afforded strong support for the hypotheses, the implications would have been readily generalized to include responses elicited by individual methods of Rorschach examination. Since the data failed to offer strong support for the hypotheses, however, the findings are regarded with caution. Analyses must be designed which can be conducted with material from individually conducted protocols. It is possible that exposing a list of alternative responses to a subject biases him (particularly if he is suggestible), and may detract from trends which could be clearly demonstrated in material from individually administered Rorschach tests. It seems unlikely, however, that well-defined patterns of response could be revealed in one testing situation and altogether obscured in another. It is suggested, therefore, that the present data may be taken as evidence that the provisions of Assumptions I and II will not be sufficiently marked in individually administered protocols to warrant the confident use which is constantly made of them.

The scant support which the Hypotheses A, B, and C receive is taken as indication that the usual abstract scoring procedures are of no value in



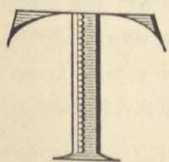
attempts to appraise the behavioral significance of Rorschach responses elicited by check-list procedures.

## REFERENCES

1. BALINSKY, B. The multiple choice group Rorschach test as a means of screening applicants for jobs. *J. Psychol.*, 1945, 19, 203-208.
2. BECK, S. J. *Rorschach's test*. New York: Grune and Stratton, 1945. Vols. I and II.
3. CHALLMAN, R. C. The validity of the Harrower-Erickson Multiple Choice test as a screening device. *J. Psychol.*, 1945, 20, 41-48.
4. CHESIRE, L., SAFFIR, M. AND THURSTONE, L. L. *Computing diagrams for the tetra-choric correlation coefficient*. Chicago: Univ. of Chicago Bookstore, 1938.
5. ENGLE, T. L. The use of the Harrower-Erickson Multiple Choice test in differentiating between well adjusted and maladjusted high school pupils. *J. Educ. Psychol.*, 1946, 37, 550-556.
6. HARROWER-ERICKSON, M. R., AND STEINER, M. E. *Large scale Rorschach techniques*. Springfield, Ill.: Charles C Thomas, 1945.
7. JENSEN, M. B. AND ROTTER, J. B. The validity of the multiple choice Rorschach test in Officer Candidate selection. *Psychol. Bull.*, 1945, 42, 182-185.
8. KLOPFER, B. AND KELLEY, D. M. *The Rorschach technique*. Yonkers, N.Y.: World Book Co., 1942.
9. LAWSHE, C. H., JR. AND FORSTER, MAX H. Studies in projective techniques: I. The reliability of a multiple choice group Rorschach test. *J. Appl. Psychol.*, 1947, 31, 199-211.
10. RAPAPORT, D. *Diagnostic psychological testing*. Chicago: Year Book Pub., 1945. Vols. I and II.
11. RORSCHACH, H. *Psychodiagnostics*. Berne: Hans Huber, 1942.
12. WINFIELD, M. C. The use of the Harrower-Erickson Multiple Choice Rorschach test with a selected group of women in military service. *J. Appl. Psychol.*, 1946, 30, 481-487.

*J. R. Wittenborn*

*A FACTOR ANALYSIS OF  
RORSCHACH SCORING  
CATEGORIES*



THE PRESENT study was undertaken for the purpose of testing three hypotheses. These hypotheses were developed during the course of a series of investigations concerned with the use of the Rorschach test.

Several of the analyses which have been published in the present series of studies (4, 5) provide evidence that the traditional distinction between Rorschach responses based primarily on the perception of human movement, and Rorschach perceptual responses based on an employment of color, is congruent with the manner in which various manifestations of these two classes of responses are intercorrelated. In this respect, at least, the universally observed scoring distinction between the movement and color responses is justifiable. The distinctions between the movement and color responses provided by the various interrelationships examined were relative and far from absolute, however. Inasmuch as there is a variety of different Rorschach determinant scores, it is possible that the distinctions

Reprinted from *J. Consult. Psychol.*, 1950, 14, 261-267, by permission of the American Psychological Association and the author.



revealed between the human movement and color scores are minor relative to the distinctions that exist between some of the other determinant scores. This, however, is not implied in the Rorschach literature. As a matter of fact, in most writings the distinction between human movement and color receives a primary emphasis. Accordingly, one would not predict that this distinction would be obscured by other marked and broadly relevant distinctions between Rorschach determinants. The magnitude of the distinction between the human movement response score and the color response score may be evaluated by comparing it with the magnitude of the distinctions which exist between other classes of response. Such a comparison is relevant to Rorschach theory and may have some practical implications.

Under certain conditions, a factor analysis of the intercorrelations among various commonly employed Rorschach scoring categories should reveal the difference between the human movement and the color responses. This possibility is stated formally in Hypothesis I.

*Hypothesis I:* In a factor analysis of the various Rorschach response categories, the pattern of factorial composition for the human movement response will be different from the pattern of factorial composition for responses involving the use of color.

In order for the factor analysis to be relevant to Hypothesis I, it is necessary for both the human movement response score and some of the color response scores to show large correlations with several scoring categories, i.e., it is necessary for both human movement and color scores to have an important amount of common factor variance. If, under this condition, the analysis does not reveal evidence for Hypothesis I, the distinctions between the human movement and the color responses which have been revealed in earlier analyses may be considered to be minor relative to the distinctions which exist between other response categories.

As a result of a few different fragmentary clues provided by earlier analyses (3, 5), it was inferred that some of the Rorschach responses differed from each other with respect to the degree of perceptual control characterizing them. Although the exact nature of this "control" was not and is not clearly specified, human movement responses and large detail responses were conceived as involving a high order of perceptual control, whereas whole responses and color-form responses were considered to be alike<sup>1</sup> in that they involve a flexible perceptual approach requiring appreciably less

1. To what degree the suspected behavioral similarity between color-form and whole responses is due to a possible prevalence of whole color-form responses to one or two cards is unknown, but the concept of uncontrolled perceptual responses as expressed in Hypothesis II is extended beyond such specific possibilities and the concept must stand or fall on its broad implications.

perceptual discipline. A low order of perceptual control is tentatively conceived to be manifested by responses which are incautious, possibly spontaneous or impulsive, to a degree which results in a relative disregard for or an unawareness of the purely formal, literal, or concrete response possibilities. In order to apply this conceptualization broadly, the pure color, the pure texture, and the pure diffusion response categories may be added to the whole and the color-form response categories. Similarly, in order to broaden the tentatively conceived class of responses which expresses a high order of perceptual control, the pure form responses and the other detail response categories are added to the human movement and the large detail response categories. This conceptualization requires that the pattern of intercorrelation among the response categories should be organized in the manner specified by Hypothesis II.

*Hypothesis II:* The whole, color-form, pure color, pure texture, and pure diffusion response categories are intercorrelated in a manner which yields a factor different from the factor which is most important in determining the common factor variance of the pure form, human movement, and the various detail response categories.

A classification of responses on the basis of the degree to which they require perceptual control is closely akin to the common practice of evaluating responses on the basis of the degree to which they reflect spatially definite well-conceived formal percepts. The conceptualization responsible for Hypothesis II is somewhat different from the usual form-level evaluation in that it is intended to emphasize distinctions in perceptual approach rather than distinctions in perceptual quality or plausibility. It also makes no provision for the qualitative distinctions among the texture, diffusion, and color determinants. It implies that the distinctions between responses which are based upon the degree of perceptual control transcend the distinctions among the different classes of determinants. Hypothesis II, if a true hypothesis, may invite a shift in emphasis from the qualitative distinctions between certain classes of determinants to a system of distinctions based on perceptual control; it does not preclude, however, that there are discernible systematic distinctions among whole, uncontrolled color, uncontrolled texture, and uncontrolled diffusion responses.

Both as a result of clinical applications of the Rorschach method and researches concerning the psychological significance of Rorschach responses, the writer has come to suspect that although the general productivity of the patient greatly influences the score for most response categories, it does not influence the score for the various response categories equally. If productivity is an important determiner of the frequency with which responses fall in



the various response categories, the possibility exists that most of the reliable variance of some of the scoring categories is due primarily to the general productivity of the patients. For the purposes of the present experiment, this possibility is formally stated by Hypothesis III.

*Hypothesis III:* The Rorschach scoring categories will be correlated with the total number of responses (*R*) in such a manner as to result in a general or quasi-general factor of productivity.

If Hypothesis III is supported by the present data, it becomes possible to examine the validity of Hypothesis IV which has practical implication for the use of the Rorschach test.

*Hypothesis IV:* Response-scoring categories differ greatly in the degree to which their common-factor variance is due to a productivity factor.

Hypothesis IV is worth examining because the degree to which the various scoring categories are independent of general productivity is unknown, and if the various scoring categories differ markedly in this respect, it would suggest, all other things being equal, that the scoring categories most independent of productivity be accorded a more uniquely significant meaning than those which reflect primarily the productivity of the patient, i.e., the recognition of response categories which are primarily measures of productivity would contribute to efficient and economical scoring procedures and

TABLE 1  
Definitions of Scoring Categories

Variable No.	Symbol	Definition
1	W	Whole blot
2	D	Large usual detail
3	d	Small usual detail
4	Dd	Unusual detail
5	S	White space
6	M	Figures in human-like action (human, mythological, or animal)
7	FM	Animals in animal-like action
8	m	Abstract or inanimate movement
9	K	Shading as diffusion (smoke, clouds)
10	FK	Shading as three-dimensional expanse in vista or perspective
11	F	Form only, not enlivened
12	Fc	Shading as surface appearance or texture, differentiated
13	c	Shading as texture, undifferentiated
14	C'	Achromatic surface color
15	FC	Definite form with bright color
16	CF	Bright color with indefinite form
17	C	Color only
18	P	Popular responses
19	O	Original responses, found not more than once in one hundred records
20	R	Total number of responses

TABLE 2  
Intercorrelations Among Scoring Categories

No.	Symbol	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1	W																			
2	D	-194																		
3	d	-228	669																	
4	Dd	-138	608	749																
5	S	-143	464	512	660															
6	M	-041	498	560	665	616														
7	FM	-058	625	429	381	288	359													
8	M	115	403	351	297	287	379	375												
9	K	255	330	246	259	155	269	281	425											
10	FK	203	275	238	221	232	154	201	240	411										
11	F	-052	693	705	762	635	438	285	202	147	198									
12	Fc	100	587	583	605	519	579	326	450	349	153	460								
13	c	144	138	116	108	097	113	136	182	310	193	108	191							
14	C'	016	466	510	503	508	532	345	468	310	114	377	546	076						
15	FC	023	541	429	457	374	342	361	563	181	028	305	571	-084	464					
16	CF	410	241	090	114	-008	-108	-025	113	353	285	179	117	251	089	133				
17	C	170	178	072	137	066	-020	075	098	259	270	039	098	261	109	177	235			
18	P	-095	208	266	163	034	177	221	363	198	-005	025	357	187	206	356	-033	216		
19	O	-011	436	570	579	565	624	303	456	305	135	415	646	181	379	550	046	172	659	
20	R	093	879	783	813	598	629	581	478	429	379	797	710	195	551	584	315	224	188	551



reduce unrecognized sources of error necessarily inherent in practices which involve the multiple employment of the same variable which unrecognized appears in different guises.

### *Subjects and Procedures*

As a means of testing the three hypotheses, twenty-one different Rorschach scores were intercorrelated for a sample of ninety-two Yale undergraduates and the intercorrelations were submitted to a centroid analysis. The Rorschach scoring categories employed are shown in Table 1, and their intercorrelations in Table 2. The variables may be recognized as comprising the principal scoring categories of the Klopfer system. The intercorrelations were computed by the Pearson product-moment method, which requires that the variables be linearly related to each other. An examination of numerous scatter diagrams between a variety of scoring categories gave no indication that the relationships were nonlinear. All of the distributions of the scoring categories were positively skewed, however. Had the skewness in some instances been negative, it is possible that the linearity of some of the interrelationships among the scoring categories would have been distorted. Some of the scores occurred infrequently and were not found in

TABLE 3  
Centroid Factors

No.	I	II	III	IV	h <sup>2</sup>
1	087	472	134	-417	.422
2	788	-291	159	094	.740
3	745	-353	085	184	.721
4	773	-378	122	079	.762
5	629	-341	-049	-038	.516
6	656	-301	-191	094	.566
7	540	-127	-013	196	.346
8	601	151	-296	-122	.487
9	521	409	101	053	.452
10	383	277	323	096	.337
11	664	-385	350	-071	.717
12	765	-098	-166	-120	.637
13	284	375	144	197	.333
14	649	-122	-225	-137	.505
15	613	-102	-308	-289	.565
16	283	400	461	-274	.528
17	274	358	148	110	.237
18	384	261	-456	308	.518
19	744	032	466	109	.784
20	941	-216	277	-064	1.013

every protocol. It is not customary to apply the product-moment coefficient to variables having such a small range<sup>2</sup>; nevertheless, it is felt that a system of intercorrelations such as provided by Table 2 permits an evaluation of the hypotheses in question.

The four factors yielded by the centroid analysis are presented in Table 3. The factors in Table 3 may be rotated to the position described in Table 5 by an employment of the transformation matrix shown in Table 4. All of

TABLE 4  
Transformation Matrix\*

	I	II	III	IV
I'	556	104	-721	-403
II'	195	586	-197	761
III'	316	635	537	-425
IV'	744	-494	349	281

\* In one of the rotations .709 was erroneously substituted for .707 in the sine and cosine. For this reason the rotated factors are not perfectly orthogonal (the departure from orthogonality is quite minute) and the communalities in Table 5 do not correspond perfectly with those of Table 3.

the various rotations were selected by an application of the usual criteria, i.e., maximizing the number of zero loadings and minimizing the negative loadings. The test clusters responsible for rotated factors II', III', and IV' are relatively independent of each other. The clustering tendency responsible for factor I' is not wholly distinct from that responsible for factor IV'. If the present study were concerned with factors per se, it would be necessary to rotate factors I' and IV' obliquely with respect to each other and it would be possible to make other minor improvements in the rotated factors.

From an examination of Table 5 it may be seen that, with certain qualifications, Hypothesis I is a true hypothesis. Variable 6, the human movement variable, draws most of its common factor variance from the factor IV'; it also draws a substantial quantity of its factor variance from factor I'. The factorial composition of variable 6 is quite different from that of variables 16 and 17, the color-form and the pure color response variables; most

2. In general it may be observed that the variables that have the smallest range tend to have relatively small correlations. This is not surprising, but this difficulty cannot be resolved in any simple manner. Since all the correlations are based upon the same sample, the usual considerations regarding heterogeneity of sample are not applicable nor is it clearly appropriate to regard the difficulty as a result of broad categories because an attempt to correct the variables in question for a broad category effect would involve ascribing statistically a refinement and sensitivity which actually does not characterize the scores. Since the present investigation is concerned with the conceptual implications of the Rorschach *scoring procedures*, it was decided to study the scores directly with whatever imperfections they may possess and not attempt statistically to refine them or their intercorrelations.



TABLE 5  
Rotated Factor Matrix

No.	Symbol	I'	II'	III'	IV'	h <sup>2</sup>
1	W	169	-015	576	-239	.420
2	D	255	023	110	812	.737
3	d	242	061	-021	810	.719
4	Dd	271	-035	036	827	.760
5	S	365	-007	-028	609	.505
6	M	433	061	-127	597	.564
7	FM	217	183	-001	515	.346
8	m	612	171	179	235	.491
9	K	238	361	456	236	.450
10	FK	-030	246	429	288	.328
11	F	105	-220	183	786	.711
12	Fc	583	033	141	526	.637
13	c	014	397	321	132	.278
14	C'	565	-005	065	426	.505
15	FC	669	-100	086	318	.566
16	CF	-024	-010	707	097	.510
17	C	039	318	347	109	.235
18	P	445	552	-089	084	.518
19	O	709	339	-041	406	.784
20	R	327	292	336	885	1.088

of their common factor variance may be explained in terms of factor III'. On the basis of the present analysis, however, the form-color response variable (number 15) is more similar in its factorial composition to the human movement variable than it is to the other color variables. The factorial composition of the form-color response is not at all similar to the factorial composition of the color-form and the pure color responses. The form-color response is different from the human movement response in that it draws most of its common factor variance from factor I' and is only secondarily dependent on factor IV', which is the chief source of the common factor variance of the human movement response variable. It is interesting to observe that the factorial composition of the form-color response is quite similar to the factorial composition of the achromatic color responses (variable 14), the form-texture responses (variable 12), and to a lesser degree the inanimate movement (variable 8) and the human movement (variable 16) responses. This pattern of relationships is scarcely predictable from the Rorschach literature and the scoring and interpretive practices of Rorschach workers. The present sample suggests that the superficial logical similarity between the form-color and the color-form responses may have little behavioral basis. If the pattern of relationships revealed in the present sample can be verified in other samples, it may be desirable to consider a revision of our

current beliefs concerning the behavioral or personality significance of the form-color response.

An examination of factor III' shows it to be the factor required by Hypothesis II. Factor III' may be seen to be determined by the classes of response (whole, pure vista, pure texture, color-form and pure color) which were judged by the writer on a priori grounds to be similar in that they presumably require a loose, informal, spontaneous perceptual approach not governed primarily by the most obvious spatial characteristics of the ink blot. Factor III' is also seen to be correlated with variable 10, the form-vista variable; although no prediction concerning variable 10 was made. Variable 20, total productivity (*R*), participates to some degree in factor III'. This is not an untoward result and obviously does not detract from the meaning of factor III'.

A most impressive feature of Table 5 is its failure to correspond with the qualitative classifications of Rorschach determinants (movement [*M*, *FM*, *m*], vista [*K*, *FK*], texture [*c*, *Fc*] and color [*C*, *CF*, *FC*]) which are emphasized so insistently in the literature and about which elaborate and tediously arbitrary scoring distinctions have been devised. For example, the factorial compositions of the form-vista, pure vista, pure texture, and pure color response variables are strikingly similar to each other and involve primarily factors II' and III'. The form-color and form-texture response variables have a highly similar factorial composition which does not involve factors II' and III', but instead draws upon factors I' and IV'. Thus, we see that qualitatively similar Rorschach response determinant variables not only fail to cluster together in a manner to form factors, but they have dissimilar factorial compositions and responses which fall in one particular qualitative determinant classification are inclined to resemble most certain responses which belong to some other class of determinants. Findings such as these suggest that the behavioral or personality significance traditionally ascribed to many of the specific Rorschach scoring categories may be incorrect in its emphasis even if valid to a degree.

Factor IV' seems to satisfy the requirements of Hypothesis III. This would be more obvious if factors I' and IV' were rotated obliquely with each other, or if the reference vector corresponding to factor IV' were projected through variable 20. Nevertheless, Hypothesis IV may be discussed in terms of factor IV'. The relative importance of general productivity may also be discerned by an examination of the intercorrelations, Table 2. For the most part, the variables which are important to factor IV' also have substantial loadings with factor I' and in general these variables have a highly similar factorial composition; specifically, all of the various detail scores are highly similar to each other and to a striking degree are similar in factorial compo-



sition to all of the scores which are based on responses which primarily emphasize form. At any rate, either the correlations with variable 20 (total number of responses, given in Table 2) or the loadings of the various scores on factor IV in which variable 20 is so important may be taken as evidence that many of the Rorschach response scores are for the most part highly similar to each other and are also a measure of the productivity of the subject in the Rorschach situation. All of the various types of scores are not equally affected by the general productivity of the patient. Some of the Rorschach scores, particularly those which have high loadings with factors II' and III', show important evidence of having a behavioral significance quite different from the behavioral significance of the general productivity of the subject.

It may be noted that variable 18, the number of popular responses, is less important in the productivity factor than variable 19, the number of original responses. This probably arises from the fact that there is a limited number of popular responses; because of their popularity the percentage of popular responses for moderately productive subjects would be as great or greater than the percentage of popular responses for highly productive subjects. Similarly, since originality of responses is based on a frequency of occurrence criterion, the subjects producing the largest number of responses, all other things equal, have a greater chance of producing original responses.

In general, the manner in which variables are intercorrelated depends upon patterns of differences among the individuals comprising the sample. The pattern of differences among the individuals comprising the present sample is in some respects reminiscent of clinically familiar personality patterns. This is particularly true in the case of the variables which cluster together in a manner to form factor III'. The tendency to give numerous whole responses and to give responses in general which are spontaneous, impressionistic, and not exclusively formal and literal in their spatial aspects has often been found to characterize the so-called hysterical personality. Such perceptual tendencies on the Rorschach are also often considered to be characteristic of the classical personality trait of extroversion. To what degree either the hysterical personality or the extrovert personality produces a relatively large number of the kinds of Rorschach responses which determine factor III' would not be difficult to determine. As a matter of fact, it would be very simple from the arithmetic standpoint to weight the types of responses contributing to factor III' in a manner to yield a scale of the response tendency or characteristic responsible for the factor. If this were done, the behavioral significance of this scale could be explored by correlating it with results or other personality tests and with different clinical and behavioral criteria.

The Rorschach responses responsible for factor IV', viewed as a group, are also reminiscent of a familiar personality reaction pattern which has considerable clinical interest. Specifically, Rorschach protocols produced by ruminative, obsessive, compulsive individuals are commonly characterized by a preponderance of detailed responses, relatively numerous human movement, animal movement and pure form responses, a high level of productivity and a tendency to avoid informal responses such as those contributing to factor III'.

The participation of *W* responses in factor III' may be puzzling to some readers and they may be inclined to suspect that its presence in factor III' is due to a tendency for *K*, *c*, *CF*, and *C* to be determinants for whole responses. To what degree this is true cannot be established from the present analysis but in an earlier study (3) several whole responses to achromatic cards were found to be highly correlated with the *CF* responses to chromatic cards 8 and 9.

### *Conclusions*

As a means of evaluating certain hypotheses concerning the behavioral similarities and dissimilarities of Rorschach responses, twenty-one of the basic scores of the Klopfer system were intercorrelated and submitted to a centroid analysis. The intercorrelations could be accounted for in four factors and after a series of orthogonal rotations several clustering tendencies became conspicuous. As a result of this analysis and within the limitations of the sample of ninety-two students employed, the following conclusions are offered:

1. The factorial composition of the human movement response category is distinctly different from the factorial composition of the color-form and the pure color response categories. The color-form and the pure color responses have a highly similar factorial composition.

This result combined with the results of earlier studies (4, 5) comprises considerable justification for the traditional distinction (between human movement and the color responses) which is emphasized in the scoring of the Rorschach test and in an interpretation of Rorschach protocols.

2. The factorial composition of the form-color response category is quite different from the factorial composition of the color-form and color response categories. This result suggests that the common practice of regarding the three classes of color responses as similar in their implications for the affective features of human behavior should be viewed with reservations. In the present study, the factorial composition of the form-color response cate-



gory is more similar to the factorial composition of the human movement response category than it is to the other color response categories.

3. Factor III' in the present study fulfills the requirement of Hypothesis II which states, "The whole, color-form, pure color, pure texture, and pure diffusion response categories are intercorrelated in a manner which yields a factor different from the factor which is most important in determining the common factor variance of the pure form, human movement, and the various detail response categories." The evidence for the validity of Hypothesis II suggests that it may be profitable to employ a new Rorschach scoring principle which is based on a weighted combination of the response categories which contribute to factor III'.

4. Factor IV' may be offered in support for both Hypothesis III ("The Rorschach scoring categories will be correlated with the total number of responses [*R*] in such a manner as to result in a general or quasi general factor of productivity") and Hypothesis IV ("Response scoring categories differ greatly in the degree to which their common factor variance is due to a productivity factor"). There are two possible practical implications for factor IV':

a) Factor IV' may be produced by the obsessive, compulsive personalities in the sample and when its contributing response categories are appropriately weighted, it may provide a scale which has some clinical value.

b) This factor has implications for the current use of the Rorschach test inasmuch as it implies that some of the separately interpreted scoring categories overlap each other and should receive relatively less weight than they now do.

5. In general, the present study suggests that an incorrect emphasis may have influenced the development of current Rorschach scoring procedures and interpretive practices. For example, many of the scoring categories which belong to the various broad classes of determinants, e.g., color, texture, or diffusion have a quite dissimilar factorial composition and in general the manner in which the various determinant scoring categories cluster together could not be predicted by an employment of the usual beliefs concerning behavioral implications of determinants.

## REFERENCES

1. KLOPPER, B. AND KELLEY, D. M. *The Rorschach technique*. Yonkers, N.Y.: World Book Co., 1942.
2. WITTENBORN, J. R. Certain Rorschach response categories and mental abilities. *J. Appl. Psychol.*, 1949, 33, 330-338.

3. WITTENBORN, J. R. A factor analysis of discrete responses to the Rorschach ink blots. *J. Consult. Psychol.*, 1949, 13, 335-340.
4. WITTENBORN, J. R. Statistical tests of certain Rorschach assumptions: Analyses of discrete responses. *J. Consult. Psychol.*, 1949, 13, 257-267.
5. WITTENBORN, J. R. Statistical tests of certain Rorschach assumptions. The internal consistency of scoring categories. *J. Consult. Psychol.*, 1950, 14, 1-19.



J. R. Wittenborn

STATISTICAL TESTS OF  
CERTAIN RORSCHACH  
ASSUMPTIONS • *The Internal*  
*Consistency of Scoring Categories*

IN THE first publication of this series, *Analyses of Discrete Responses* (13), two reasons were given for examining the implications of some of the Rorschach assumptions:

1. Since the assumptions have not been tested, the possibility exists that some of them are invalid. If some of the assumptions cannot be validated, aspects of interpretations based on the questionable assumptions should be employed with reservation or perhaps eliminated from standard procedures.

2. Many of the assumptions employed in Rorschach procedures cannot be found among the established facts and theories of American academic psychologists. If the Rorschach assumptions are valid, their relevance to the

Reprinted from *J. Consult. Psychol.*, 1950, 14, 1-19, by permission of the American Psychological Association and the author.

psychology of personality organization and to the psychology of perception could be of extraordinary significance.

In scoring a Rorschach protocol, it is customary to classify each response with respect to several classes of factors, such as content, portion of ink blot interpreted, and ink-blot characteristics determining the response. There are standard categories within each of the classes of factors. When a response is scored, the content category in which it falls, as well as its location category and its determinant category, is designated by the scorer. As a second step in scoring, the total number of responses falling in each category is ascertained. Regardless of the precise manner in which the categories are used, the device of categorizing the responses and then finding the total number of responses in each category carries two implications:

I. That all of the responses falling in a given category are similar in some behavioral respect.

II. That the psychological significance of responses falling in a given category is different in some respect from responses placed in other categories.

These assumptions are implicit, for example, in the distinction that is made between human movement and color responses. This distinction was first emphasized by Rorschach (9), and its observance has remained a basic feature of Rorschach psychodiagnosis. Because of the basic significance ascribed to these two types of responses, it was decided to employ this distinction in testing the foregoing assumptions of intracategory similarity and intercategory difference.

If Assumption I is a true proposition, it follows that several measurements of the tendency to perceive human movement would be positively related. If several measures of a given response category were not positively interrelated, it would be difficult to justify an assumption that differences in number of responses of that category expressed differences in a personality attribute. It also follows from Assumption II that several measures of a given response category would be more highly interrelated with each other than they would be related with measures of some other category.

The following experimental hypotheses have been tested by statistical analyses (A and B) of data provided by the Harrower-Erickson multiple choice check list (3), the analyses were based on a sample of 247 undergraduates:

A. The tendency for controlled color responses (*FC* and *CF*) to be associated with other controlled color responses (*FC* and *CF*) is greater than the tendency for such color responses to be associated with human movement (*M*) responses.



B. The tendency for human movement (*M*) responses to be associated with other human movement (*M*) responses is greater than the tendency for such responses to be associated with controlled color responses (*FC* and *CF*).

C. The degree to which Rorschach responses will be intercorrelated is a function of the number of scoring categories the responses have in common.

The major conclusions of the Analyses (A and B) of the check list data may be stated as follows (13):

1. Among the intercorrelations between pairs of discrete check list responses there is no obvious pattern which conforms with a pattern predictable from either Hypothesis A or Hypothesis B.

2. There is a statistically significant tendency for Rorschach responses which have two or three scoring categories in common to be more highly correlated than pairs of responses which have but one or no common scoring categories. Thus Hypothesis C is verified. Nevertheless, the tendency for the correlations between responses to vary directly with the number of common scoring categories is so negligible as to be associated with a large number of gross exceptions.

3. In general, the number and size of the exceptional correlations is such as to challenge the value of scoring the Rorschach responses on the basis of abstract categories.

As a result of these findings, it was decided to submit the discrete check list responses to a factor analysis (Analysis C) which could provide an answer to the following questions:

1. The intercorrelations among the responses could not be satisfactorily predicted by certain deductions from Assumption I and II: is there some structure, some systematic clustering effect among the check list responses or are their relationships haphazard and unpatterned?

2. If there is a discernible pattern of relationships among the check list responses, will it be congruent with any current practice or belief concerning the Rorschach?

From the factor analysis (14), which involved 18 responses and was based upon 247 students, the following conclusions may be offered:

1. The analysis yielded relatively independent factors, most of which retained their composition when rotated obliquely. Nevertheless, an inspection of the factors yielded by both orthogonal and oblique methods of rotation does not offer evidence in support of some of the common beliefs concerning the Rorschach responses, particularly those which are reflected in abstract scoring practices.

2. The data suggest that response content (with its implied associations and projections) may play an important role in determining the functional

similarities and dissimilarities of certain responses, particularly those which involve movement.

Because of its known practical limitations, a Rorschach multiple choice check list per se is of restricted interest; nevertheless, the form lends itself to quantitative analyses which in turn yield hypotheses relevant to and readily tested by the individual form of the test. With perhaps some exceptions it may reasonably be supposed that the responses which the individual has checked on the check list comprise a sample of the perceptions which may be elicited from him by the ink blots. Therefore, in the absence of specific contra-indications, it is justifiable to generalize the results of certain statistical analyses (which are convenient only with data yielded by the check list form of the test) to relevant assumptions which have implications for all forms of the test. Consequently, inferences from the preliminary studies have been employed in the formulation of hypotheses which are tested in the present study with data yielded by the individual Rorschach procedure.

### *The Plan of the Present Investigation*

In the present study three different analyses are employed for the purpose of testing eleven different hypotheses, almost all of which were generated primarily by Assumptions I and II. In order to facilitate the presentation and to clarify the nature of the study each of the three analyses will be presented separately with a formal statement of the hypotheses to be tested, a description of the subjects employed, a terse presentation of the statistical test results, and a brief discussion of the apparent status of the hypotheses.

#### ANALYSIS D

In order to test the consequences of Assumptions I and II, it is necessary to secure several different measures or quantifiable groupings of the scoring categories in question. This has been accomplished in Analysis D by determining for each card the number of responses falling in a given scoring category. Thus each card may be considered to yield its own evidence of the individual's tendency to give responses which belong to a given scoring category.

A measure of a response tendency based on a single card involves certain considerations which should be made explicit:

1. Such measures of a response tendency are but a fraction of the total protocol score for that response tendency. Accordingly, the reliability of such a measure



cannot be expected to approach the reliability of the total score for the response tendency (i.e., scoring category) in question.

2. Moreover, such measures of a response tendency cannot be expected to be equally reliable from card to card. Therefore, perfect evidence for the assumptions of intracategory similarity and intercategory difference should not be expected.

3. Although the frequency of responses of a given category is known to vary conspicuously from card to card, this is not expected to affect the tests of the internal consistency (i.e., the correlation between two variables is independent of any constant difference between them).

4. Regardless of the exact manner in which responses of a given category are fractionated, the resulting subgroups will differ from each other in some qualitative respect. These qualitative differences will obviously detract from the evidence for intracategory similarity. Some readers may be inclined to seize upon such qualitative differences and argue that they disqualify a fractional analytic examination of Rorschach assumptions. It is to be emphasized, however, that such qualitative differences are the prime justification for studies of internal consistency such as the present one; if such qualitative differences are so great as to obscure completely the trends required by the hypothesis, then the trends specified in the hypothesis may be rejected as irrelevant and perhaps the nature of the qualitative differences may be considered to be the proper subject for analysis.

Analysis D is designed to test two hypotheses:

D. There will be more significant interrelations *among* the number of human movement responses for cards I, II, III, VII, IX, and X than there will be *between* the number of human movement responses for these same cards (I, II, III, VII, IX, and X) and the number of color responses (total of *FC*, *CF*, and *C*) for cards II, III, VIII, IX, and X.

E. There will be more significant interrelationships *among* the number of color responses (total of *FC*, *CF*, and *C*) for cards II, III, VIII, IX, and X than there will be *between* the number of human movement responses for cards I, II, III, VII, IX and X and the number of color responses (*FC*, *CF*, and *C*) for cards II, III, VIII, IX, and X.

Hypotheses D and E were tested by means of data provided by the following three samples:

Sample I—Comprises a group of ninety-five Yale undergraduates who voluntarily served as subjects for a Ph.D. candidate, Mrs. Florence Schumer, who administered and scored the Rorschach tests in a manner closely conforming to that of Klopfer (4).

Sample II—Comprises forty-five Yale undergraduates who had consulted the writer for guidance and therapy. All could be described as moderately neurotic but none was incapacitated by his symptoms nor in apparent danger of entering a psychotic episode. The Rorschach test protocols employed were prepared by Klopfer's methods.

Sample III—Comprises a group of 100 patients who had been treated on either an out-patient or an in-patient basis in the clinics of the psychiatric service of the New Haven Hospital. Almost all of these patients could be described as ill,

and at least one-third were psychotic or had suffered a psychotic episode. Most of the examiners were close adherents to the Klopfer method, although several of them were equally familiar with the methods of Beck. (The scores employed in the analysis were determined in conformance with Klopfer.)

It is always possible that findings based on a particular sample, regardless of its composition, will be considered by some readers as inappropriate for testing such an instrument as the Rorschach. Basing analyses on several groups which are clearly different in the quality of their personal adjustment affords an important control.

The data for each sample were of such a nature as to make the following statistical tests feasible:

Sample I—55  $\chi^2$  tests were made comprising the following groups:

1. 15  $\chi^2$  tests, based on all possible interrelationships among the number of human movement responses for cards I, II, III, VII, IX and X.
2. 10  $\chi^2$  tests based on all possible interrelationships among the number of color responses for each of the colored cards, II, III, VIII, IX, and X.
3. 30  $\chi^2$  tests based on all possible interrelationships between the number of human movement responses for cards I, II, III, VII, IX, and X and the number of color responses for cards II, III, VIII, IX, and X.

Sample II—45  $\chi^2$  tests were made comprising the following groups:

1. 10  $\chi^2$  tests based on all possible relationships between the number of human movement responses for cards I, III, VII, IX, and X. (Because of the small size of sample B an analysis of the human movement responses for card II was not made.)
2. 10  $\chi^2$  tests based on all possible interrelationships among the total number of color responses for each of the colored cards, II, III, VIII, IX, and X.
3. 25  $\chi^2$  tests based on all possible interrelationships between the number of human movement responses for cards I, III, VII, IX, and X, and the number of color responses for cards II, III, VIII, IX, and X.

Sample III—55  $\chi^2$  tests were made comprising groups of  $\chi^2$  tests which were arranged identically with those provided by sample I.

The 155  $\chi^2$  tests provided by the analysis of samples I, II, and III are summarized in Table 1 below. These data have the following implications:

1. Since almost half (eighteen out of forty) of the interrelationships among groups of human movement responses would occur by chance less than 10 per cent of the time, Hypothesis D is considered to be a true hypothesis. That is to say, there is a discernible degree of mutual consistency among human movement responses, and as a consequence it is inferred that the human movement response category is a behaviorally relevant category in the sense that it refers to demonstrably similar ways of reacting to ink-blot stimuli. Since human movement responses comprise functionally similar behavioral features, quantification of them is an appropriate procedure. The consistency among groups of human movement responses (as well as



TABLE 1

A Summary of the Results of Chi-Square Tests Required by Hypotheses "D" and "E,"  
Showing the Distribution of Significance Levels for All Groups Compared

Sample	<i>Significance level distribution of the comparison for each of the three possible combinations of the determinant groups</i>		
	M vs. M	M vs. C	C vs. C
Sample I			
1% level	3	1	1
5% level	6	5	4
10% level	6	6	6
99% level			
(All comparisons for Sample I)	15	30	10
Sample II			
1% level	2	0	6
5% level	3	0	8
10% level	4	0	8
99% level			
(All comparisons for Sample II)	10	25	10
Sample III			
1% level	4	1	2
5% level	6	1	5
10% level	8	2	6
99% level			
(All comparisons for Sample III)	15	30	10
All Samples Combined			
1% level	9	2	9
5% level	15	6	17
10% level	18	8	20
99% level			
(All comparisons for all samples)	40	85	30
Approximate fraction significant at 10% level	$\frac{1}{2}$	$\frac{1}{10}$	$\frac{2}{3}$

their relative independence from groups of color responses) may be taken as evidence that the total human movement response score could bear a valid relationship to an important feature of the personality which could not be predicted from a knowledge of the individual's color responses.

2. The similarity among small groups of human movement responses is not of a high order, and it is not invariably consistent. Since exceptions to the hypothesis exist, the common clinical practice of ascribing a particular

meaning (or a particular personality implication) to small differences in the number of human movement responses is obviously hazardous and probably unjustified. That is to say, small groups of human movement responses do not invariably have the same meaning or the same behavioral implications; it is gratuitous to ascribe any particular difference between two individuals as a consequence of observed small differences (one or two responses) between them with respect to the number of human movement responses they produce.

3. Hypothesis E is also a true hypotheses. Two-thirds of the interrelationships among groups of color responses are so marked that they would occur less than 10 per cent of the time by chance (i.e., significant at the 10 per cent level). Since less than one-tenth of interrelationships between groups of human movement responses and groups of color responses are significant at as high a level, the practice of combining the number of color responses into a total score which is interpreted differently from the human movement total score appears to be justified. Moreover, since the color scores are related with each other, it is quite possible that the total color response scores could bear an important degree of relationship with some other response score, e.g., a measure of some practically important feature of personality. It is evident from the data that small groups of responses to the color cards may not always have the same meaning. It is interesting to note, however, that the consistency with which a particular meaning may be ascribed to responses to the different color cards appears to be substantially greater than the consistency of meaning which may be ascribed to the human movement responses elicited by different cards.

4. The status of Hypotheses D and E does not appear to be dependent upon the exact nature of the sample employed inasmuch as the findings for all three samples appear to have the same implications for the two hypotheses.

5. There is no evidence that the pattern of significant  $\chi^2$  tests is consistent from sample to sample. Accordingly, it may not be concluded that there is a greater similarity among the human responses (or among the color responses) for any particular group of cards than there is for any other group of cards. This lack of consistency from sample to sample is in keeping with the interpretation that the number of insignificant  $\chi^2$  tests is an expression of a generally low level of intercard relationships among color or among human movement determinant scores.

#### ANALYSIS E

In Analysis D hypotheses generated by the assumptions of intracategory similarity and intercategory difference among abstract Rorschach scoring



categories were examined for two determinants, human movement and color. The hypotheses, D and E, required a demonstrable degree of similarity among groups of responses which had a common determinant but which were different with respect to the card which elicited them. For Analysis D the consequences of the assumption of similarity among responses with a common determinant were tested by means of an experimental design wherein differences due to the location of the responses were distributed among the groups in an uncontrolled manner. Such differences obviously reduce the sensitivity of the analysis, but they probably do not result in im-

TABLE 2

The Confidence Levels (Based on the Chi-Square Test) at Which the Statistical Hypothesis of "No Relationship" May be Refuted for Various Pairs of Response Grouping

Deter- minant	Loca- tion	Sam- ple	<i>M</i>			<i>FC</i>			<i>CF</i>		
			D	W	X	D	W	X	D	W	X
M	D	I									
		III									
	W	I	1%								
		III	5%								
FC	X	I	1%	—							
		III	10%	5%							
	D	I	(5%)*	—	—						
		III	(—)	—	10%						
	W	I	—	(—)	—	(5%)					
		III	—	(—)	5%	—					
CF	X	I	10%	—	(5%)	10%	—				
		III	—	—	(1%)	—	5%				
	D	I	(5%)	—	—	(—)	—	10%			
		III	(—)	—	10%	(5%)	10%	5%			
	W	I	—	(10%)	—	—	(1%)	—		10%	
		III	—	(—)	—	—	(—)	—		—	
	X	I	—	—	(—)	—	—	(10%)		—	—
		III	—	—	(5%)	—	5%	(1%)		5%	—

\* Some of the confidence levels are enclosed in parentheses. The comparisons to which they refer are between groups with the same location. They are employed in Analysis F and not in Analysis E.

portant systematic differences between the groups which were being tested. Since the response groups employed in Analysis D were determined by the cards, differences due to the card (e.g., content differences resulting from the nature of the stimulus material) vary systematically from group to group and to some degree detract from the apparent validity of the hypotheses.

Analysis E is designed somewhat differently. Like Analysis D, E is concerned with testing hypotheses which state the consequences of Assumptions

I and II with respect to the human movement and color determinants. In Analysis E, however, intercategory differences due to content (due to card) are uncontrolled, and may randomly reduce the sensitivity of the tests of relationship whereas the groups differ systematically with respect to location. Accordingly, differences due to the location factor systematically reduce the demonstrable magnitude of the relationship between the groups. In Analysis E three location groupings are employed: (1) The *W* group comprises all of the human movement or all of the color responses in the protocol which are scored as whole responses. (2) The *D* group comprises all the movement or color responses of the protocol which are scored as based upon one of the large details of the card. (3) The *X* group comprises all other movement or color responses which were not scored as whole responses or as large detail responses; for the most part the *X* group comprises responses based upon small or rarely used portions of the card. Since fewer groups are employed in Analysis E than in Analysis D, in Analysis E a larger number of responses is available for each group. As a consequence, it is possible for the color responses to be subdivided so that the provisions of Assumptions I and II may be examined as they apply to the degree to which color responses are controlled by form. Analysis E makes use of the data provided by samples I and III described in Analysis D. The data from these two samples were employed independently in testing the following hypotheses.

Hypothesis F—Groups of human movement responses which are different from each other with respect to their location (*W*, *D*, or *X*) are more consistently correlated with each other than they are with groups of color responses (*FC* and *CF*) which are different from the groups of human movement responses not only with respect to determinant but also are different from the human movement responses with respect to location.

Hypothesis G—Groups of color responses (*FC* and *CF*) which are different from each other with respect to their location (*W*, *D*, or *X*) are more consistently correlated with each other than they are with groups of human movement responses which are different from the groups of color responses not only with respect to determinant but also are different from the color responses with respect to location.

Hypothesis H—Groups of *FC* responses which are different from each other with respect to location are more consistently correlated with each other than they are with groups of *CF* responses which are different from the *FC* responses with respect to location as well as with respect to the degree to which the response is controlled by form (*FC* and *CF* are here employed to represent the responses which they commonly designate in the Klopfer scoring system).

Hypothesis I—Groups of *CF* responses which are different from each other with respect to location are more consistently correlated with each other than they are with groups of *FC* responses which are different from the *CF* responses with respect to location as well as with respect to the degree to which the response is controlled by form (*CF* and *FC* are here employed to represent the responses which they commonly designate in the Klopfer scoring system).



The  $\chi^2$  tests required by hypotheses F, G, H, and I are summarized in Table 2. In the instance where the relationship between the groups of responses is so small as to occur more often than 10 per cent of the time by chance the confidence level is omitted from the table; these cases are indicated by a dash. Since the data for samples I and III were analyzed independently, the results for each sample are shown. The significance of the relationship between any two particular groups may be readily found in the Table by locating the row for one of the groups in question and the column for the other group; the point at which the respective row and column intersect gives the significance of the relationship between the two groups.

Confidence levels in Table 2 which are relevant to Hypothesis F are summarized in Table 3. Since five of the six possible interrelationships among the groups of human movement responses (three comparisons for each of the two samples involved) are significant at the 10 per cent confidence level and since only four of the twenty-four possible interrelationships between human movement and color response groups are equally significant, Hypothesis F is considered to be a true hypothesis. The evidence for Hypothesis F is noticeably more consistent than the evidence in favor of the analo-

TABLE 3

A Summary of the Results of the Chi-Square Tests Required by Hypothesis "F," Showing the Distribution of Significance Levels for the Groups Compared

	M vs. M	M vs. C
Samples I and III		
1% level	2	0
5% level	4	1
10% level	5	4
99% level (All location groups combined)	6	24
Approximate fraction significant at 10% level	$\frac{5}{6}$	$\frac{1}{6}$

gous hypothesis, D. Some of this difference may be due to the greater reliability inherent in the larger groups employed in Hypothesis F. It seems more probable, however, that the difference is a result of the difference between the designs employed in Analyses D and E. The support provided Hypothesis F favors the practice of using the total number of human movement responses as a quantification for a particular response mode. The total number of human movement responses appears to be an internally consistent feature of behavior in the Rorschach situation and accordingly could possess an important degree of validity for some personality criterion.

The confidence levels relevant to the status of Hypothesis G are similarly

summarized in Table 4. It is apparent that significant relationships between two groups of responses which involve color are twice as prevalent as they are between two groups of responses which do not have the same determinant. It is interesting to note that despite the fact that the groups employed in Hypothesis G are presumably more reliable than those employed in Hypothesis E, the evidence for Hypothesis G is not as conspicuous as the evidence for the analogous Hypothesis E. The difference, probably a feature of the design of the analyses, will be discussed in a succeeding section. The

TABLE 4

A Summary of the Results of the Chi-Square Tests Required by Hypothesis "G," Showing the Distribution of Significance Levels for the Groups Compared

	C vs. C	M vs. C
Samples I and III		
1% level	0	0
5% level	5	1
10% level	9	4
99% level (All location groups combined)	24	24
Approximate fraction significant at 10% level	$\frac{1}{3}$	$\frac{1}{6}$

evidence for Hypothesis G provided by the present analysis supports the practice of combining the color responses into a total score and suggests that the tendency to respond to color in Rorschach situations represents a consistent human trait or behavior attribute which is different from the tendency to offer human movement responses.

TABLE 5

A Summary of the Results of Chi-Square Tests Required by Hypothesis "H," Showing the Distribution of Significance Levels for the Groups Compared

	FC vs. FC	FC vs. CF
Samples I and III		
1% level	0	0
5% level	2	2
10% level	3	4
99% level (All location groups combined)	6	12
Approximate fraction significant at 10% level	$\frac{1}{2}$	$\frac{1}{3}$

The summarized material in Table 5 suggests that Hypothesis H could be a true hypothesis but the distinction is not clear; there is no evidence among the data provided by Table 2 to support Hypothesis I. Nevertheless, these data are not considered as a challenge to the practice of distinguishing between the *CF* and the *FC* score.



The rather modest evidence in Analysis E for consistency among the color responses is in contrast with the fact that in Analysis D two-thirds of the interrelationships among the color responses were found to be significant. This contrast is interpreted as a consequence of the difference between the two experimental designs and will receive a more nearly complete discussion in a subsequent section.

#### ANALYSIS F

Analysis F like Analysis E employs samples I and III which were described and employed in Analysis D. In other respects, however, Analysis F is greatly different from Analyses D and E. Analyses D and E were concerned with hypotheses which specified consequences of Assumptions I and II for the Rorschach scoring distinction between responses involving human movement and those involving color. Analysis F is concerned with hypotheses which have to do with distinctions in Rorschach scoring which are based upon the portion of the card presumably eliciting the response. In Analysis E the portions of the card eliciting the response (commonly referred to in scoring as the location factor) were described with respect to three subdivisions. These subdivisions are employed in the present analysis also and comprise the following groups:

- I. The number of responses based on the whole card, *W*.
  - (a) *W* responses with the *M* determinant.
  - (b) *W* responses with the *CF* determinant.
  - (c) *W* responses with the *CF* determinant.
- II. The number of responses based on the large details of the card, *D*.
  - (a) *D* responses with the *M* determinant.
  - (b) *D* responses with the *FC* determinant.
  - (c) *D* responses with the *CF* determinant.
- III. The number of responses which have a location score other than *W* or *D*; in the present study the location of these responses is designated by the letter *X*.
  - (a) *X* responses with the *M* determinant.
  - (b) *X* responses with the *FC* determinant.
  - (c) *X* responses with the *CF* determinant.

In the case of Hypothesis K differences due to card and to content are uncontrolled, but differences due to the determinants vary systematically from group to group and systematically detract from the apparent magnitude of the relationship between the groups.

Hypothesis K—Response groups which belong to the same scoring category with respect to location (but are different with respect to determinant) are more consistently related among themselves than they are with response

groups which belong to a different location category (and are different with respect to determinant).

Confidence levels for the relationships relevant to Hypothesis K may be found in Table 2 and for the convenience of the reader are summarized in Table 6. On the basis of the present data, Hypothesis K is considered to be a true hypothesis. Nevertheless, only one-half of the interrelationships among groups of responses which have the same location but a different determinant are significant at the 10 per cent level. The evidence of functional similarity among groups of responses which have the same location is of academic interest and is in keeping with the practice of counting and interpreting separately those responses which are elicited by different portions of the card. Rorschach examiners commonly ascribe a personality difference between individuals as a result of small observed differences with respect to the number of whole or large detail responses; the data and the analyses of this study may *not* be taken as a justification for this practice. As a matter of fact, they may better be taken as cause for viewing such practice with suspicion.

TABLE 6

Levels of Significance of Relationship Distributed for Response Groups Which Have the Same Location and Distributed for Response Groups Which Have a Different Location

	Significance of Relationships Between Groups Having the Same Location	Significance of Relationships Between Groups Having a Different Location
1% level	1	0
5% level	5	1
10% level	6	4
99% level (All comparisons)	12	24
Approximate fraction significant at 10% level	$\frac{1}{2}$	$\frac{1}{6}$

The *FC* and *CF* categories, although scored differently, were not shown in Analysis E to be conspicuously or consistently different from each other. In Rorschach theory they are considered to be expressions of the same general class of response and are commonly combined to yield a total color response score. In keeping with the conception of the *FC* and *CF* as merely gradations within the response to color category they have been combined in a preliminary analysis (8) as well as in Analysis D of the present series. Viewing the *FC* and *CF* responses as having a determinant in common results in the statement of Hypothesis L which like the foregoing hypotheses



is generated primarily from Assumptions I and II. For the purpose of testing Hypothesis L groups of responses which have *FC* as their determinant are considered to have the same determinant as groups of responses for which the determinant is actually *CF*. Thus it is possible to interpret certain of the parenthetical values in Table 2 as confidence levels for relationships between response groups which are the same with respect to both determinant and location features.

**Hypothesis L**—Responses which have both a common location and a common determinant category are more consistently related with each other than are responses which have either a location or a determinant factor in common.

Hypothesis L has an interesting relevance to Assumptions I and II because it states that similarities among responses which are due either to common location or determinant categories are cumulative, and in this respect it is similar in its implications to Hypothesis C, which was tested in an earlier study and is stated in the introduction to the present paper.

The significance levels relevant to Hypothesis L are distributed in Table 7. The significance levels distributed in the first row of Table 7 are based upon relationships between response groups which are identical with respect to location and are similar with respect to determinant, i.e., the groups compared all have color as a determinant but the groups compared are different

TABLE 7  
Distribution of Significance Levels

	1% Level	5% Level	10% Level	99% Level (All Com- parisons)	Approximate Fraction Significant At 10% Level
Relationships between groups of responses which have a common location and a quasi-common determinant	2	3	4	6	$\frac{2}{3}$
Relationships between groups of responses which have a location only in common	1	5	6	12	$\frac{1}{2}$
Relationships between groups of responses which have a common determinant, either quasi or strictly defined	2	9	13	30	$\frac{1}{2}$
Relationships between groups of responses which have neither a location nor a determinant category in common	0	1	4	24	$\frac{1}{6}$

with respect to the degree to which color is controlled by form. The second row in Table 7 comprises a distribution of significance levels determined between groups of responses which are identical with respect to location only. The third row comprises a distribution of significance levels between groups of responses which are different from each other with respect to location and are identical with respect to determinant (groups which are strictly identical with respect to determinant as well as those which have a quasi-common determinant are both included). The last row in Table 7 is a distribution of significance levels for the relationships between groups of responses which have neither a location nor a determinant category in common. It may be seen from an examination of Table 7 that the distributions of levels of significance are arranged in a manner in keeping with the provisions of Hypothesis L.

#### ANALYSIS G

Analysis G employs data provided by samples I and III. The hypotheses tested in Analysis D are homogeneous with respect to design and implication, as were those tested in Analyses E and F. In the present analysis the hypotheses tested, M and N, have their origins in preceding studies and in Rapaport (8) and Rorschach (9).

The Rorschach literature does not state explicitly whether location factors or determinant factors are more important in the abstract scoring of Rorschach responses, nor is there any evidence in the literature which could be taken as a definite indication that determinant factors are more (or less) consistent in influencing the psychological significance of the response than are the location factors. In the factor analysis of discrete responses to the Rorschach check list (8) no evidence was produced to indicate that determinant factors played a more important role in influencing the interrelationships among the responses than did the location factors. The relevance of this question to the kind of data employed in the present analysis is stated in Hypothesis M.

Hypothesis M—The portion of significant relationships among groups of responses which have a common determinant but differ with respect to location is as great as the portion of significant relations between groups of responses which have the same location but are different with respect to determinant.

Table 8 provides a summary of the significance levels which are relevant to Hypothesis M. It is apparent that the data available for the present analysis do not challenge the status of the hypothesis. In general it appears that the consistency among groups of responses which have a common determi-



TABLE 8  
Distribution of Significance Levels

	1% Level	5% Level	10% Level	99% Level (All Com- parisons)	Approximate Fraction Significant At 10% Level
Relationships between groups of responses which have a common location (D, W, or X)	1	5	6	12	$\frac{1}{2}$
Relationships between groups of responses which have a common determinant (M, FC, or CF)	2	7	10	18	$\frac{1}{2}$

nant is of about the same order as that among groups of responses which have a common location. It should be emphasized, however, that the number and diversity of comparisons involved in the present analysis is not sufficient to confer any general validity to Hypothesis M; it can only be stated that Hypothesis M is not challenged by the available data.

It is a common belief among workers who have had intensive experience with the Rorschach that the personality significance of the human movement response varies somewhat with the exact nature of the human movement which is perceived. It is also commonly observed among Rorschach workers that the cards differ with respect to the kinds of human movement responses they elicit (5, 6). These common informal beliefs based on the clinical experience of Rorschach workers are in good agreement with the results of the factor analysis of discrete responses (14). If these beliefs growing out of clinical practice as well as the implications of the factor analysis are valid, it would be predicted that:

Hypothesis N—The interrelationships among groups of human movement responses which are based on different cards would be less marked and less consistent than interrelationships among groups of human movement responses which are not based on different cards and which do not maximize this source of differences in content.

In the case of the large detail responses and the whole responses, one would not often expect the nature of the perceived human movement to be contingent upon the subject's willingness to use the whole card in preference to merely a large portion of it in forming his response. In general it seems plausible to suppose that the location scoring designation of a response has less to do with its content than the card which elicited the response. This seems particularly plausible inasmuch as the exact location designation of a human movement response seems to be more often deter-

mined by combinatory or relatively incidental features of the response than by the exact nature of the *human movement* feature of the response. The foregoing considerations are of interest because they predict the shift in the prevalence of significant relationships between human movement response groups observed upon comparison of Table 1 (Analysis D) with Table 2 (Analysis E). Hypothesis N is considered to be a true hypothesis and its status and implications are further discussed in the following section.

## Discussion

The nature of the procedure employed in testing the hypotheses requires comment. It is to be noted that all of the hypotheses require a comparison of two samples of interrelationships. The interrelationships comprising each sample were based upon groups of responses which were not independent of each other; they instead were related to each other to an unknown but varying degree. Estimation of the sampling distributions of the  $\chi^2$  statistic for such samples of interrelationships would be inconvenient, if not prohibitively difficult. Accordingly, the chance frequency of any particular difference between the sample distributions of the  $\chi^2$  statistics is unknown, and it is not possible to specify the confidence level with which the null statement of the hypotheses may be refuted. In the case of each of the hypotheses, it is possible to state only whether the difference between the distributions of the  $\chi^2$  statistic for the two samples of relationships is in a direction predictable by the hypothesis. Regardless of the subjective confidence which the reader may have in the status of any of the hypotheses, it should be noted that the evidence is favorable to all the hypotheses except possibly those involving the distinction between the *FC* and *CF* responses. Accordingly, the acceptance of Assumptions I and II as true propositions (within the requirements of the stated hypotheses and the characteristics of the Rorschach samples employed) should not be regarded as reckless or incautious.<sup>1</sup>

In evaluating the present analyses and their implications for Rorschach practices, it should be noted that despite the relatively large samples employed, a large fraction of the relationships required by the hypotheses are insignificant. Since there is no apparent pattern which leads to a distinction between the significant and insignificant  $\chi^2$  tests for a given sample of relationships, the relationships exceptional to the requirements of the hypotheses are considered to be a result of either the unreliability of the response groups or of a modest degree of intrinsic relationship among the responses

1. The assumptions do not require that the Rorschach scoring categories are optimal for classifying responses and the present study is not designed to indicate the nature of the optimal or most economical scoring categories.



which have a common abstract scoring category. In any event, the present data are in no sense a justification for the common practice of ascribing important personality differences between individuals as a result of small differences in the frequency with which color responses, movement responses, whole responses, or large detailed responses appear in the Rorschach protocols of the individuals. How large such differences must be before they may be justifiably used in the evaluation of individuals, if they may ever be so used, cannot be precisely inferred from the present analyses. Obviously, important differences between people cannot be predicted on the bases of one or two abstract response scores. (It is conceivable, however, that the *content* of a single response or the conditions under which it is elicited, may have an important validity.) Thus with respect to the first question which prompted this series of studies, it may be stated that some important provisions of Assumptions I and II afford a partial prediction of the pattern of interrelationships among the Rorschach responses. This may be taken as an indication favorable to some of the abstract scoring practices, but the data must also be taken as distinctly unfavorable to certain common clinical practices involving these abstract response scores.

The results of the analyses are particularly relevant to the second general question instigating these studies. The manner in which responses having a common abstract scoring category tend to cluster together, and tend to distinguish themselves from responses having other types of abstract scoring features, suggests that the modes of perceptual response to which some of the abstract scoring features refer may be members of a fairly broad and stable class of perceptual response modes, and may actually be evidences of practically significant classes of human attributes (relatively stable response patterns). Such response patterns, like most response patterns, would be a reflection of certain important features of the conditions of the individual's existence. If the total constellation of such response patterns were known, abstract classes of perceptual responses could possibly be of value in inferring some of the important conditions of the individual's prior existence and of perhaps greater value in predicting the conditions under which certain types of behavior could be elicited in the future.

Evidence for the assumptions of intracategory similarity and intercategory difference among responses with a common scoring designation is not of academic interest only. The evidence for the assumptions not only hints at the existence of unexplored perceptual response attributes but more specifically is an indication that the total score for some of the response categories could be of valid means of predicting a personality characteristic. As a matter of fact the present series of studies is in certain respects a validation of the Rorschach test. The present studies have sought evidence of the

validity of the Rorschach by means of tests of internal consistency. In situations where no dependable, quantified criteria are available for validating a test, the possibility of a test's being valid may be explored by examining its internal consistency. If the numerous component features of a test are found to be unrelated with each other, the total score of the test cannot be validly related to any simple criterion.

In a preceding paragraph, it was suggested that the tendency to perceive human movement in response to the Rorschach cards could comprise a portion of a broad sample of response patterns which were not restricted to the Rorschach situation. If the human movement response is so conceived, the nature of the general class of responses under which it may be subsumed becomes of interest. Speculation concerning such a class of responses may be guided somewhat by consideration of the conditions which determine or qualify the human movement response. In clinical practice it is commonly observed that the presence or absence of human movement responses having a particular content is determined by fantasies, conflicts or repressions of the individual (7). As a matter of fact, the relationship between the content of the human movement perceived and the content of the individual's fantasies is considered by many clinicians to be conspicuous. In the factor analysis of discrete Rorschach responses there was some evidence that the interrelationships among responses, including those involving human movement, were determined in part by content. In the present paper, there is an important difference between Tables 1 and 2 with respect to the prevalence of significant relationships among groups of human movement responses. Only half of the interrelationships among human movement response groups which differed from each other with respect to card were significant at the 10 per cent level whereas all but one of the interrelationships among human movement response groups which differed from each other with respect to location were significant at the 10 per cent level of confidence. This shift may be interpreted as an expression of the importance of content in determining the human movement response. This interpretation emerges from these considerations: By virtue of their form the cards differ importantly among themselves with respect to the kinds of human movement responses readily elicited; that is to say, the human movement content which the form of the card favors appears to determine the frequency with which certain individuals perceive human movement in the respective card; such differences due to card content are obscured, however, in interrelationships among response groups which differ systematically with respect to location, but do not differ systematically with respect to card (content).

Thus it appears that individual differences in the readiness with which a human movement is perceived is a function of the nature of the stimulus



material. Inasmuch as there is a discernible correspondence between the objective form of the card and the human movement responses elicited, it is suggested that people differ in the readiness with which they perceive different kinds of human behavior.

The tendency of the individual to see or not to see certain kinds of human acts or attitudes in the Rorschach cards may be an instance of his readiness to see or not to see evidence for these human acts or attitudes in other situations. If this general pattern of speculation is correct, and it appears to be congruent with Rorschach's interpretation of the human movement response (9), the following assumptions may be regarded as tentative indications of relationships between the Rorschach human movement perceptual response tendency and perceptual tendencies characteristic of the individual in other situations requiring responses involving human acts or attitudes:

III. Any predominant content among the human movement responses elicited by the Rorschach cards should be demonstrably related to the content of human behavior perceptions elicited by situations wherein varied alternative perceptions are possible.

IV. A person who produces a relatively large number of human movement responses on the Rorschach will tend in other situations (wherein varied alternative perceptions are possible) to produce a relatively high number of responses involving human acts, attitudes or attributes.

These assumptions are not offered as clinical aids; as a matter of fact, much clinical practice presupposes their validity. Their value lies in the possibility that upon definition of terms they may be interpreted in the form of hypotheses which specify experimental procedures.

A thorough study of the human movement responses would involve devising a wide variety of tests which sample separate aspects or phases of the broad hypothetical class of responses which include, in addition to human movement responses, other types of responses, e.g., the expression of suppressed or repressed fantasy material. These tests would be included in a battery which would also include tests for other human attributes which might be confused with the hypothetical classification of responses. If the hypothetical class of responses is properly conceived and adequately sampled by tests, a factor analysis of the intercorrelations among the tests should provide evidence relative to the status of the Assumptions III and IV.

A similar design could possibly provide an advantageous means for examining the nature and implications of responses which are based on color. The present series of investigations appears to offer few hints concerning the nature of the color responses. A few tangential observations may be made, however. It should be noted for example, that color responses in Table 1 (Analysis D) are more consistently related with each other than those given in Table 2 (Analysis E). It appears that the differences in con-

tent due to the nature of the card do not detract from the interrelationships among the color responses whereas differences due to the amount of the card involved (i.e., location) do detract from the interrelationships among color responses. This suggests that the implications of color responses may be more a function of the size of the area of the card involved in the response than the nature (content) of the response. It is of interest to observe in the factor analysis (14) that the *FC* response appeared in the same factor with responses based upon large details whereas *CF* responses occurred in factors involving the whole card responses.

It would be in keeping with a portion of the meaning generally ascribed to the color responses to hypothesize a continuum which describes the degree to which perceptions vary from the extreme of rigid control and formalization (e.g., simple form responses) through various levels of control (e.g., *FC* and *CF*) to perceptions which are uncontrolled by and incongruent with formal considerations (e.g., *C*). Because of their very nature, relatively uncontrolled responses (*CF*) may be inconsistent and less highly related with each other than the relatively controlled responses (*FC* and *F*). The idea of a hypothetical continuum of perceptual control expressed in the Rorschach cards by the degree to which color responses are controlled by form is important in the Rorschach method and it is of interest to note that, by an employment of a very tenuous reasoning, correspondence between the present findings and this important assumption is discernible. For example in Table 2, it may be seen that the *FC* responses are more highly related with the *M* response (*M* responses are highly dependent on form) than are the *CF* responses. Obviously, the present data are inadequate to the task of even tentatively defining the color response, but it is of interest that an interpretation of the present findings which is in keeping with the control concept of the color responses can be construed. If Rorschach responses based on small or rare details are considered to be most controlled and whole responses are considered to be least controlled, additional evidence for this hypothetical continuum can be construed from some other results; e.g., it may be observed in Table 2 that the whole responses show the smallest number of consistent interrelationships whereas responses involving the *X* location (made up principally of small details) shows the largest number of significant relationships. It should be observed, however, that the slight difference in the number of significant relationships to which this paragraph refers may well be due to chance.

Regardless of the validity of the foregoing conjectures, if degree of control in perceptual responses is conceived as a human attribute, it should be possible after definition of the word "control" to formulate a variety of quantifiable instances of this perceptual attribute (instances from the Rorschach as well as from other situations) and study the relative consistency of



their interrelations in the presence of measures of other attributes, the nature of which might be confused with the hypothetical control attribute. Factor analysis offers a suitable statistical approach for such a study.

At present neither the meaning nor the practical import of the human movement response or of the color response is objectified. Authoritative statements concerning them do not fully agree, and such statements comprise terms which for the most part are obscure or circularly defined. This probably is largely due to the fact that few unambiguous quantitative criteria for different aspects of the personality are available (quantified criteria for some Rorschach scores have been reported [2] but as far as the writer is aware are not described). Since such objective criteria for defining the meaning of the Rorschach responses are lacking, their meaning is not only unspecified but must vary from psychologist to psychologist and from day to day. (The uncanny competence of some clinicians is no more relevant to this problem than the incompetence of others.)

This state of affairs is important, not only from the standpoint of psychology as a behavioral science or from the standpoint of the problems in teaching it creates, but (since all of the subjective concepts of the implications of human movement and of color responses cannot be equally valid) it is important from the standpoint of the patient's welfare.

Research oriented toward the validation of these responses has produced promising but unsatisfactory evidence of validity. Since Rorschach validation research has had the handicap of questionable criteria, the following approaches to this important problem may be suggested:

1. The systematic development of objective, quantified criteria for personality characteristics important in the validation of the tests.
2. An hypothetic-deductive approach wherein an explicit unambiguous interpretation of a response group is given; numerous tests generated by this interpretation are prepared; intercorrelations of these tests are examined for mutual consistency; and the explicit interpretation rejected as invalid if the resultant tests are not mutually consistent.

The present discussion will be recognized as favoring the second of the two suggested alternatives.

### *Summary and Conclusion*

Three samples comprising a total of 240 individually administered Rorschach tests were employed in testing a series of hypotheses generated by the following assumptions:

I. That all of the responses falling in a given category are similar in some behavioral respect.

II. That the psychological significance of responses falling in a given category is different in some respect from responses placed in other categories.

As a result of the application of a variety of analytical designs the following conclusions may be offered.

1. Within the requirements of the stated hypotheses and the characteristics of the samples employed, Assumptions I and II may be accepted as true hypotheses.

2. Human movement responses comprise functionally similar behavioral elements and quantification of them is an appropriate procedure. The consistency among groups of human movement responses (as well as their relative independence from groups of color responses) may be taken as evidence that the total human movement response score could bear a valid relationship to an important feature of the personality which could not be predicted from a knowledge of the individual's color responses.

3. The practice of combining the number of color responses into a total score which is interpreted differently from the human movement total score appears to be justified. Moreover, since the color scores are related with each other, it is quite possible that the total color response scores could bear an important degree of relationship with some other response score, e.g., a measure of some practically important feature of personality.

4. The evidence of functional similarity among groups of responses which have the same location is in keeping with the practice of counting and interpreting separately those responses which are elicited by different portions of the card.

5. The present data are in no sense a justification for the common practice of ascribing important personality differences between individuals as a result of small differences in the frequency with which color responses, movement responses, whole responses or large-detail responses appear in the Rorschach protocols of the individuals. How large such differences must be before they may be justifiably used in the evaluation of individuals, if they may ever be so used, cannot be precisely inferred from the present analyses. Obviously, important differences between people cannot be predicted on the bases of one or two abstract response scores.

6. Responses which have both a common location and a common determinant factor are more consistently related with each other than are responses which have either a location or a determinant factor in common.

7. The portion of significant relationships among groups of responses which have a common determinant but differ with respect to location ap-



pears to be as great as the portion of significant relations between groups of responses which have the same location but are different with respect to determinant.

8. The data offered no strong support for the practice of distinguishing between the *CF* and the *FC* responses. The findings are not interpreted as comprising a direct challenge to this practice, however.

9. It appears that individual differences in the readiness with which a human movement is perceived is a function of the nature of the stimulus material.

10. The implications of color responses may be more a function of the size of the area of the card involved in the response than the nature (content) of the response.

11. The whole responses show the smallest number of consistent inter-relationships whereas responses involving the *X* location (made up principally of small details) shows the largest number of significant relationships.

In the discussion it is suggested that an employment of a hypothetic-deductive approach for the investigation of the validity of the Rorschach and other projective tests may at present prove to be the most feasible.

## REFERENCES

1. BECK, S. J. *Rorschach's test*. New York: Grune & Stratton, 1945. Vols. I and II.
2. GUSTAVE, A. Estimation of Rorschach scoring category by means of an objective inventory. *J. Psychol.*, 1946, 22, 253-260.
3. HARROWER-ERICKSON, MOLLY R. AND STEINER, MATILDA E. *Large scale Rorschach techniques*. Springfield, Ill.: Charles C Thomas, 1945.
4. KLOFFER, B. AND KELLEY, D. M. *The Rorschach technique*. Yonkers, N.Y.: World Book, 1942.
5. LINDNER, R. M. Some significant Rorschach responses. *J. Crim. Psychopathol.*, 1943-44, 5, 775-778.
6. LINDNER, R. M. Content analysis in Rorschach work. *Rorschach Res. Exch.*, 1946, 10, 121-129.
7. PIOTROWSKI, Z. The M, FM, and m responses as indication of changes in personality. *Rorschach Res. Exch.*, 1936-37, 1, 148-156.
8. RAPAPORT, D. *Diagnostic psychological testing*. Chicago: Year Book Publishers, 1945. Vols. I and II.
9. RORSCHACH, H. *Psychodiagnostics*. Berne: Hans Huber, 1942.
10. THORNTON, G. R. AND GUILFORD, J. P. The reliability and meaning of erlebnistypus scores in the Rorschach test. *J. Abnorm. Soc. Psychol.*, 1946, 31, 324-330.

11. WITTENBORN, J. R. Certain Rorschach response categories and mental abilities. *J. Appl. Psychol.*, 1949, 33, 330-338.
12. WITTENBORN, J. R. Statistical tests of certain Rorschach assumptions: Analyses of discrete responses. *J. Consult. Psychol.*, 1949, 13, 257-267.
13. WITTENBORN, J. R. A factor analysis of discrete responses to the Rorschach ink blots. *J. Consult. Psychol.*, 1949, 13, 335-340.
14. WITTENBORN, J. R. AND SARASON, S. B. Exceptions to certain Rorschach criteria of pathology. *J. Consult. Psychol.*, 1949, 13, 21-27.



*Irwin J. Knopf*

# RORSCHACH SUMMARY SCORES IN DIFFERENTIAL DIAGNOSIS<sup>1</sup>

**T**HE DISCREPANCY between reports of clinical experience and research data has been a serious dilemma in Rorschach circles for a number of years. On the one hand, many clinicians have been impressed with the usefulness and the validity of the Rorschach method in a range of applications, while, on the other, research findings have not in the main supported this confidence. One such application has been the considerable use of the Rorschach as an aid in differentiating psychiatric disorders.

Formulated on the assumption that differences in psychiatric conditions would be reflected in Rorschach data, early investigations were primarily concerned with descriptive reports of the typical Rorschach performance of one or more psychiatric populations. As a result, statistical and/or experimental controls were not generally employed, but instead, clinical description was accepted without more rigorous verification. Later, some workers attempted to isolate the qualitative differences purported between patient groups and yet preserve the "holistic" nature of the test findings by deriving patterns or signs which collectively seemed to discriminate among nosologi-

Reprinted from *J. Consult. Psychol.*, 1956, 20, 99-104, by permission of The American Psychological Association and the author.

1. A portion of this paper was presented at the American Psychological Association meetings in New York, 1954.

cal groups. Thus, a variety of signs were reported, for example, signs which were found in brain-damaged individuals, in psychoneurotics, in schizophrenics, and which were useful as indices of adjustment in the evaluation of psychotherapy (17, 18, 14, 8, 22, 15). However, when many of these signs were employed in subsequent investigations, the significant discriminations reported earlier were not corroborated (16, 3, 21, 7, 11). The fact that signs derived in the original investigations provided better discrimination within the initial sample than within subsequent populations is not too surprising. In some of these studies, signs were obtained by selecting for prediction those few aspects of Rorschach performance which showed the highest relationships from among the available predictors which had a low correlation with the criterion. Such a procedure tends to capitalize on chance fluctuations within the sample, and consequently may result in a spuriously high multiple-correlation coefficient. When signs which were derived in this manner are applied to subsequent populations, it can be expected that the coefficient will be of lower magnitude and of less predictive value than that obtained in the parent population.

Other investigators have studied the extent to which Rorschach single summary scores can discriminate among psychiatric groups. Wittenborn and Holzberg (23) studied thirty-nine summary scores with five patient groups and found that one score (*CF*) significantly discriminated the manic from the depressed patients. Freidman (6) evaluated the discriminative effectiveness of Rorschach scores with two groups of normal adults and 30 schizophrenics. He found eight Rorschach variables which significantly differentiated the schizophrenics from both groups of normal subjects. Reiman (19) evaluated eighty-six scores with replicated samples of ambulatory schizophrenics and neurotics. The results indicate that six scores were significant at the .10 level of confidence or better between the clinical groups for both samples. Kobler and Steil (12) reported no statistical differences of any consequence between the Rorschach scores of the paranoid and depressive subgroups of involutional melancholic patients.

While the findings with respect to single summary scores are predominantly negative, unequivocal evaluation of these data is difficult. Most studies were able to obtain a few scores which discriminated among psychiatric groups. However, in some instances, the positive findings could be expected by chance because of the large number of tests of significance computed. It should also be noted that although a variety of Rorschach scores have been reported as differentiating diagnostic groups, there has not been a great deal of consistency for these scores to appear repeatedly as diagnostic from study to study. In addition, certain methodological limitations such as small samples, incomplete statistical or experimental controls, and vaguely



defined diagnostic criteria have complicated the interpretation of the results.

In the light of the ambiguous nature of the research findings, and the extensive application of the instrument, the need for a systematic evaluation of the Rorschach as a psychodiagnostic technique seems indicated. An investigative program of this sort is under way at the Iowa Psychopathic Hospital. The present study represents one major phase of this research program, and it specifically deals with the problem of determining the extent to which Rorschach summary scores can differentiate psychiatric groups.

### *Procedure*

Our subjects were selected on the basis of the following criteria: (a) chronological age of fifteen or older; (b) unanimous agreement among psychiatrists as to diagnosis both on admission to and discharge from the hospital; (c) diagnosis was restricted to psychoneurosis (*Pn*), psychopathy (*Pp*), or schizophrenia (*Sc*); (d) diagnosis was independent of the Rorschach data,<sup>2</sup> and (e) the number of Rorschach responses (*R*) would not contribute to a significant difference in the mean number or the variance of responses for the three groups.

Initially, over 800 case records of patients who were fifteen years or older, and who were given the Rorschach test during the six-year period from 1948 to 1953, were examined in order to check on the agreement and consistency of the psychiatric diagnosis. Each case folder contained an initial diagnostic impression usually made by a psychiatric resident or a staff psychiatrist, an admission and staff diagnosis, and a final discharge staff diagnosis, both of which were made by several residents and one or more staff psychiatrists. In this way 339 Rorschach records were obtained, and the number of responses per record was determined. Statistical tests showed no significant difference between the mean number of responses for the three groups, while the variances between the groups were significantly different. Inspection of the data indicated that two cases in the *Pn* group, each with 153 responses, were contributing greatly to the heterogeneity of variance. Consequently they were withdrawn from the sample and statistical tests of means and variances were recomputed. The groups showed no differences in either the total number of responses or variances, so that the effects of *R* on other Rorschach scores for the three groups were considered approximately equal.

2. It was not possible to obtain complete independence of diagnosis and Rorschach data. However, this criterion was met to the extent that the initial impression and the admission staff diagnosis were made prior to the Rorschach administration.

TABLE 1

## Frequencies of Subclinical Types Within Each Major Diagnostic Group

Psychoneurotics		Psychopaths		Schizophrenics	
Mixed	31	Psychopathic pers.	38	Paranoid	46
Anxiety	30	Path. sexuality	18	Simple	13
Psychasthenia	28	Path. emotionality	14	Hebephrenic	8
Hysteria	22	Psychotic episodes	12	Mixed	7
Neurasthenia	5	Neurotic traits	8	Catatonic	4
Impulse neurosis	5	Inadequate type	7	Acute	3
Hypochondriasis	4	Asocial trends	3	Defect state	1
Psychosomatic	2	Alcoholism	2	Unclassified	18
Reactive depress.	2	Paranoid tendencies	2		
Unclassified	2	Exhibitionism	1		
		Malingering	1		
Total	131		106		100

A total of 337 Rorschach protocols meeting all the criteria and obtained from 131 *Pn*'s, 106 *Pp*'s, and 100 *Sc*'s comprised the basic data for this study. In order to assure equal treatment of the data for all subjects, each protocol was rescored according to the scoring system described by Hertz (9).<sup>3</sup> The incidence of clinical types which were included within each of the three major diagnostic groups is presented in Table 1. Additional subject characteristics of the three groups are given in Table 2. From this it will be noted

TABLE 2

Subject Characteristics for the Three Groups (*N* = 337)

Sex	Psychoneurotics (56 M, 75 F)		Psychopaths (79 M, 27 F)		Schizophrenics (57 M, 43 F)	
	Mean	SD	Mean	SD	Mean	SD
Age	27.3	8.6	26.9	9.8	27.7	9.6
Educ.	11.7	2.5	11.1	2.5	12.4	3.2
Length of hosp. (days)	58.1	43.7	40.5	37.5	55.8	45.3

that there was an unequal number of males and females in each group, and that this was most discrepant in the *Pp*'s. The small differences in age between the groups were not statistically reliable, although the differences in education and length of hospitalization were significant at the .01 level of confidence. The educational level of the *Sc*'s was slightly higher than that of the *Pn*'s and *Pp*'s, while the length of hospitalization was slightly shorter for the *Pp*'s than for the other two groups. Data were also available with re-

3. The author wishes to express his appreciation to Donald Spangler for rescored each Rorschach protocol.



gard to the subject's previous admission and illness. Thirty-seven per cent of the *Pn* group had been ill prior to this admission, whereas 24 per cent of the *Pp*'s and 23 per cent of the *Sc*'s had been ill previously. Generally these figures indicate that approximately 72 per cent of the total sample were first admissions at the time of the Rorschach administration, and that their condition was not, for the most part, considered chronic.

Medians, means, and standard deviations were computed for each clinical group on the following summary scores: *R*, *W*%, *D*%, *Dr*%, *S*, *F*%, *F* + % of 80, *F* + % of 60, *M*, *FM*, *Fm*, *FC*, *CF*, *C*, *Fch*, *chF*, *ch*, *Fch'*, *ch'F*, *ch'*, *Fc*, *cF*, *c*, *A*%, *Ad*, *H*%, *Hd*, Nature, Blood, Sex, Anatomy, Object, Fire, Position, Contamination, *P*, *O*%, and Rejection. Inspection of these figures indicated marked differences in the medians and the means for many of the scores, supporting previous observations that most Rorschach scores are not distributed normally (4, 5). In addition almost half of the scoring categories had medians of zero and means of less than 1.0, which not only suggests that these scores occurred very rarely, but also that they have very limited utility in individual differential diagnosis. With the exception of the *c*, *ch'*, Position, and Contamination scores which occurred very infrequently, chi-square tests of independence were employed to evaluate the significance of differences between the groups for the remaining thirty-four summary scores. The over-all median derived from the total sample (337) for each score was used as the empirical cutoff point. Yates's correction for continuity was applied to the data wherever cell frequencies were lower than ten (13). The null hypothesis was retained with those chi-square values which did not meet the minimum requirement of the .05 level of confidence.

## *Results*

Although tests of significance were not computed, the incidence of occurrence of contaminated and position responses will be presented because some Rorschach workers have regarded these responses as diagnostically important in that they are almost always associated with schizophrenia or psychosis (1, 2, 10, 20). Our data, however, indicate that these responses can and do occur in other psychiatric conditions. For example, we found contaminated responses in nine *Sc*'s, two *Pn*'s, and one *Pp*, and position responses in two *Sc*'s, five *Pn*'s, and one *Pp*. Moreover, when we consider that only twelve patients out of the total sample of 337 produced contaminated responses and only eight patients produced position responses, it is apparent that diagnostic classification cannot be effectively made solely on the basis of the presence or absence of these responses.

TABLE 3  
Medians, Means, and Sigmas for the Larger Clinical Groups on the Significant Scores  
and the Probability Values Obtained from Intergroup Comparisons

Score	$Pn (N = 131)$			$Pp (N = 106)$			$Sc (N = 100)$			Intergroup chi-square value probabilities*		
	Mdn	Mean	SD	Mdn	Mean	SD	Mdn	Mean	SD	$Pn-Pp$	$Pp-Sc$	$Pn-Sc$
<i>Dr%</i>	11.0	14.4	13.8	6.5	10.0	11.4	15.0	15.8	12.8		.01	
<i>A%</i>	50.0	48.0	17.4	51.0	50.2	18.0	44.0	43.4	17.0		.01	
<i>FM</i>	2.0	2.7	2.4	2.0	2.6	2.5	2.0	2.0	2.0			.02
<i>P</i>	5.0	5.1	2.2	5.0	5.2	2.5	4.0	4.1	2.0		.01	.02
<i>Sex</i>	0.0	0.5	1.2	0.0	0.2	1.0	0.0	0.9	2.8	.01	.01	
<i>Anatomy</i>	1.0	2.5	3.3	1.0	1.7	2.4	1.0	2.6	4.3	.01		.01

\* Only probability values at or beyond the .05 levels are reported.



The over-all chi-square tests applied to each Rorschach score for the three clinical groups resulted in significant values for the following six scores: *Dr%*, *A%*, *FM*, *P*, *Sex*, and *Anatomy*. Three additional chi-square values were computed for each of these significant scores to determine more specifically which group or groups the scores discriminated among. Table 3 lists the medians, means, and standard deviations for these scores, as well as the probability level obtained from the separate comparisons of the clinical groups. Most apparent from this table is the failure of any one score to discriminate among all three groups, although three scores were discriminative in two comparisons. It will also be noted that five scores significantly differentiated the *Pp*'s from the *Sc*'s, whereas only two scores differentiated the *Pn*'s from the *Pp*'s, and the *Pn*'s from the *Sc*'s. Inspection of the data and reference to the medians and means listed in Table 3 indicates the direction of significance. There were more *Pp*'s who were lower on *Dr%*, and higher on *FM* than *Sc*'s; more *Pn*'s who were lower on *Dr%*, and higher on *A%* than *Sc*'s; more *Pn*'s who were higher on *FM* than *Sc*'s; more *Pp*'s and *Pn*'s who were higher on *P* than *Sc*'s; and more *Pn*'s and *Sc*'s who were higher on *Sex* and *Anatomy* responses than *Pp*'s.

The *F + %* score is often regarded as important in differentiating psychotic from non-psychotic subjects. However, the over-all median value of 80 per cent which was obtained from the total sample did not discriminate among the clinical groups. Beck (2) has suggested 60 per cent as a diagnostically useful cutoff point, and consequently this score was employed with the present data. The chi-square values indicated significant differences between the *Sc*'s and the *Pn*'s, and the *Pp*'s and *Pn*'s at the .001 and .05 levels of confidence respectively, while no significant difference was obtained between the *Pp*'s and the *Sc*'s. Recognizing that there are differences in the Hertz and Beck scoring systems which include *F + %* tables and procedures for computing *F + %*, these findings nevertheless suggest that Beck's empirical cutoff point of 60 per cent may also be discriminative with Hertz's scoring method.

In evaluating the results, it seemed important to consider the extent to which chance factors could account for the significance of differences between the groups for the seven Rorschach scores. Having computed thirty-four over-all chi-square values, we can expect approximately two to be significant merely by chance at the .05 level of confidence. In the light of this possibility, a more rigorous examination of the stability of the present findings seemed essential. Therefore, 150 cases, fifty from each clinical group, were randomly selected from the total parent sample of 337. Statistical treatment and the analysis of the data for this sample was the same as that previously described for the parent population. Chi-square values were ob-

TABLE 4  
Medians, Means, and Sigmas for the Clinical Groups of 50 Cases on the Significant Scores  
and the Probability Values Obtained from Intergroup Comparisons

Score	$P_n$ ( $N=50$ )			$Pp$ ( $N=50$ )			$Sc$ ( $N=50$ )			Intergroup chi-square value probabilities*		
	Mdn	Mean	SD	Mdn	Mean	SD	Mdn	Mean	SD	$Pn-Pp$	$Pp-Sc$	$Pn-Sc$
<i>Dr%</i>	8.0	12.7	13.0	7.0	9.9	11.1	14.0	16.5	13.2		.01	.05
<i>M</i>	1.0	1.8	2.2	1.0	1.8	1.5	1.0	1.6	2.8		.01	
<i>Fch'</i>	1.0	1.1	1.5	0.0	0.5	1.3	0.0	0.5	1.0	.01		.05
<i>P</i>	5.0	4.9	2.2	5.0	5.2	2.4	4.0	4.0	2.0		.01	
<i>Sex</i>	0.0	0.6	1.6	0.0	0.1	0.4	0.0	1.0	2.5		.02	
<i>Anatomy</i>	1.0	2.6	3.2	1.0	1.5	1.9	2.0	2.8	2.9	.05	.01	

\* Only probability values at or beyond the .05 levels are reported.



tained separately for the thirty-four Rorschach scores, and only those scores which were significant on both samples were regarded as stable.

The analysis of the data for the new sample revealed that the *Dr%*, *M*, *Fch'*, *P*, *Sex*, and *Anatomy* scores significantly differentiated the groups at the .05 level of confidence or better. The medians, means, and standard deviations for each score together with the probability levels obtained from the separate comparisons of the groups are shown in Table 4. These results indicate the *Pp*'s and the *Sc*'s were significantly differentiated by five scores, whereas only two scores differentiated the *Pn*'s from the *Pp*'s, and the *Pn*'s from the *Sc*'s. In comparing these findings with those obtained from the parent sample, we find, that only the *Dr%*, *P*, *Sex*, and *Anatomy* scores significantly discriminated the clinical groups on both samples, and thus met the criterion of stability. While all four stable scores discriminated the *Pp*'s from the *Sc*'s, these scores were less sensitive in differentiating the *Pn*'s from the *Sc*'s, and the *Pn*'s from the *Pp*'s in that only one score for each comparison was found to be significant with both samples (*Dr%* and *Sex*, respectively).

### *Summary and Conclusions*

To determine the extent to which Rorschach summary scores could discriminate among psychiatric populations, a total of 337 carefully selected Rorschach records obtained from 131 *Pn*'s, 106 *Pp*'s, and 100 *Sc*'s were analyzed. Chi-square tests of independence were computed on thirty-four Rorschach summary scores for the total sample, and also for a second sample of 150 cases drawn randomly from the parent sample. Only scores which were discriminative on both samples were considered stable. The results showed that:

1. Most Rorschach summary scores are not normally distributed.
2. Almost half of the scoring categories had medians of zero and means of less than 1.0, not only indicating the rareness of these responses but also underscoring the limited utility of these scores for differential diagnosis.
3. Contaminated and position responses can and do occur, albeit infrequently, in all three groups and cannot be regarded as pathognomonic of psychosis or schizophrenia.
4. On an over-all basis, four scores: *Dr%*, *P*, *Sex*, and *Anatomy* significantly discriminated among the groups on both samples at or beyond the .05 level of confidence.
5. When specific tests of significance were made, no single summary score significantly differentiated all three clinical groups.

6. For practical purposes, Rorschach summary scores cannot be regarded as effective in differentiating psychiatric groups.

## REFERENCES

1. BECK, S. J. *Rorschach's test: I. Basic processes*. (2nd ed.) New York: Grune & Stratton, 1950.
2. BECK, S. J. *Rorschach's test: II. A variety of personality pictures*. New York: Grune & Stratton, 1947.
3. CRONBACH, L. J. Statistical methods applied to Rorschach scores: A review. *Psychol. Bull.*, 1949, 46, 393-431.
4. BERKOWITZ, M., & LEVINE, J. Rorschach scoring categories as diagnostic "signs." *J. Consult. Psychol.*, 1953, 17, 110-112.
5. FISKE, D. W., & BAUGHMAN, E. E. Relationships between Rorschach scoring categories and the total number of responses. *J. Abnorm. Soc. Psychol.*, 1953, 48, 25-32.
6. FREIDMAN, H. A comparison of a group of hebephrenic and catatonic schizophrenics with 2 groups of normal adults by means of certain variables of the Rorschach test. *J. Proj. Tech.*, 1952, 16, 352-360.
7. HAMLIN, R. M., ALBEE, G. W., & LELAND, E. M. Objective Rorschach "signs" for groups of normal, maladjusted, and neuropsychiatric subjects. *J. Consult. Psychol.*, 1950, 14, 276-282.
8. HARROWER-ERICKSON, MOLLY R. The values and limitations of the so-called "neurotic signs." *Rorschach Res. Exch.*, 1942, 6, 109-114.
9. HERTZ, MARGUERITE R. Scoring the Rorschach ink-blot test. *J. Genet. Psychol.*, 1938, 52, 15-64.
10. KLOPPER, B., & KELLEY, D. M. *The Rorschach technique*. Yonkers, N.Y.: World Book Co., 1946.
11. KNOFF, I. J. The Rorschach test and psychotherapy. *Amer. J. Orthopsychiat.*, 1956, 26.
12. KOBLER, F. J., & STEIL, AGNES. The use of the Rorschach in involutional melancholia. *J. Consult. Psychol.*, 1953, 17, 365-370.
13. LEWIS, D., & BURKE, C. J. The use and misuse of the chi-square test. *Psychol. Bull.*, 1949, 46, 433-489.
14. MIALE, FLORENCE, & HARROWER-ERICKSON, MOLLY R. Personality structure in psychoneuroses. *Rorschach Res. Exch.*, 1940, 4, 71-74.
15. MUENCH, G. A. An evaluation of nondirective psychotherapy by means of the Rorschach and other indices. *Appl. Psychol. Monogr.*, 1947, No. 13.
16. NADEL, A. B. A qualitative analysis of behavior following cerebral lesions. *Arch. Psychol.*, N.Y., 1938, 32, No. 224.
17. PIOTROWSKI, Z. A. On the Rorschach method and its application in organic disturbances of the central nervous system. *Rorschach Res. Exch.*, 1936-37, 1, 23-40.



18. PIOTROWSKI, Z. A. The Rorschach ink-blot method in organic disturbances of the central nervous system. *J. Nerv. Ment. Dis.*, 1937, 86, 525-537.
19. REIMAN, G. W. The effectiveness of Rorschach elements in the discrimination between neurotic and ambulatory schizophrenics. *J. Consult. Psychol.*, 1953, 17, 25-31.
20. RORSCHACH, H. *Psychodiagnostics*. New York: Grune & Stratton, 1951.
21. RUBIN, H., & LONSTEIN, M. A cross validation of suggested Rorschach patterns associated with schizophrenia. *J. Consult. Psychol.*, 1953, 17, 371-372.
22. THIESEN, J. W. A pattern analysis of structural characteristics of the Rorschach test in schizophrenia. *J. Consult. Psychol.*, 1952, 16, 365-370.
23. WITTENBORN, J. R., & HOLZBERG, J. D. The Rorschach and descriptive diagnosis. *J. Consult. Psychol.*, 1951, 15, 460-463.

# IV

## *Validity*



Leonard I. Schneider

## RORSCHACH VALIDATION

### *Some Methodological Aspects*

#### *Introduction*

**I**T is generally agreed in clinical circles that the Rorschach psychodiagnostic method is the most comprehensive device in the clinician's repertoire for ascertaining personality patterns. However, it is also agreed that results are highly dependent upon the skill of the interpreter. This is due to the fact that the relationships between the data yielded by the test and personality variables have not been clearly ascertained and stated. Consequently, the user of the method must rely upon a body of guesses as to the relationships involved—including those which have accumulated in the literature, plus his own experience with the test in particular, and knowledge of normal and abnormal personality in general.

Considerable space in psychological literature was at one time devoted to the question of whether or not such a state of affairs was desirable. More recently the majority of those concerned with use of the instrument have supported the position that the Rorschach needs validation (1, 2, 6, 8, 15, 16, 17, 21, 26, 30, 32, 35, 36, 39). However, the variety of opinions on the requirements for Rorschach validation are as numerous as are the theoretical and methodological positions in the entire field of psychology—as the following quotations will demonstrate.

Reprinted from *Psychol. Bull.*, 1950, 47, 493-508 by permission of the American Psychological Association and the author.

Harrison (15) claims that rigid validation procedures may be applicable to cognitive aspects of the personality but not the affective aspects. He maintains that "there must be sacrifice of objectivity and spurious precision with more reliance being put upon insight and ingenuity in the forging of tools for personality investigation." Clinical validation, he claims, is ample evidence for the present, though controlled experimentation may in the future produce more definite knowledge. Rapaport (26) and Brosin and Fromm (6) call for experimental investigation of "what form level means" by use of Gestalt experiments of form genesis and prägnance. Brosin and Fromm state further that ". . . the question of what may constitute the basis for correspondence between the responses and the personality of the subject; and the nature of the equivalence between the perceptual patterns and the general personality patterns, seem to obtain the most fruitful suggestions from Gestalt Psychology."

Sargent (32) claims that "factors taken out of context have little meaning. For the reason that it is not the absolute amount of one determinant but its relation to the whole pattern which gives it significance in the individual protocol." Thus, for her, validity studies can proceed along two lines: (1) correspondence of Rorschach impression of the total personality with impressions independently arrived at, and (2) predictive capacity which she thinks is the sounder approach due to the fact that correlations based upon inter-judge agreement are minimized by "each analyst understanding correctly different aspects of the personalities under inspection."

Macfarlane (21) lists five validation methods: (1) correlation with outside criteria, (2) projective data vs. case history data, (3) through time consistencies, (4) collateral experimental approach, and (5) degree of success in prediction, which she offers as the most fruitful method. In support of this method she says:

The writer's opinion is that the utilization of the interpretation and predictive judgments of widely experienced clinicians later checked for predictive success, will offer at this stage the most productive leads. If an experienced clinician is able to predict with considerable success, then the data on which he bases his correct intuitive predictions can be inspected and validly weighted configurations can be established, quantified, and made available to less experienced people. Also, his wrong and partially right predictions in conjunction with the right ones can serve for finer and more differentiating criteria.

Frank (8) holds that validation must be concerned with the aggregate of part functions since he believes that there may be lawfulness at that level but not between the discrete part functions and outside criteria. Like Harrison, he appears to believe in the acceptability of intermediate gross validation, for he says:



If it appears that the subject projects similar patterns or configurations upon widely different materials and reveals in his life history the sequence of experiences that make these projections psychologically meaningful for his personality, then the procedures may be judged sufficiently valid to warrant further experimentation and refinement.

Thurstone (35), in a recent polemic against the scientific laxity of Rorschach workers, calls for validation data that will enable a Rorschach interpreter to predict the style of response an individual is likely to give to different types of life situations. Hertz (16, 17) has been quite outspoken about the need for validating data. She suggests four methods: (1) direct experimentation, (2) comparison with extensive individual case studies, (3) comparison with independent, objective data, and (4) comparison with known diagnoses and clinical pictures. In her earlier article (16), she states that concentration on one individual in the tradition of the clinician would be scientifically productive.

Varvel (36) has offered some specific suggestions for validation procedures. He mentions the possibility of finding the relationship between the autokinetic phenomenon and Rorschach personality types, checking Rorschach factors against data obtained from the Dembo situation, use of drugs and post-hypnotic suggestion to measure the effects on Rorschach protocols of these experimentally produced states, and simultaneous recording with the Luria technique, pneumograph and Rorschach responses.

In 1935 Beck (1) wrote that the value of the Rorschach method was not on the basis of successful penetration into personality as yet, but as a suggestion that human nature could be passed through a prism and analyzed into component elements. He cautioned, however, that it must be recognized, ". . . that the same trait is not necessarily equivalent in two different kinds of personalities, or in another way, the same diagnostic index may have different diagnostic value depending on the larger background of the personality in which it appears." In 1942 (2) he wrote that Klopfer's argument that standardization of the Rorschach implied devising of numerical formulae for the several whole personality entities, was a fallacy. He pointed out that two issues were confused in Klopfer's position. One was the personality as a whole governed by the laws which unit personalities follow and the second was the component elements, ". . . whether we are contemplating the psychological traits or the impersonal Rorschach factors which stand for these traits." Of the latter, Beck claimed that they can and must be isolated and subjected to experimentally controlled observation independently of the problems relating to the whole personality. "The criteria of the two are different and validation is within two totally different spheres of reference."

*Point of View*

As is obvious from this sampling, a position on the most fruitful approach to take in future Rorschach validation work is dependent upon one's theoretical and methodological biases in the broader problems within the fields of science and psychology. Accordingly, before embarking on a set of criticisms and suggestions, it is necessary in the interest of clarity to make an explicit statement of position on the crucial issues involved.

Validity is here considered to be measured by a statement of the degree of concomitant variation between two independent variables, one of which is designated as the criterion variable. The data obtained from a validity investigation indicate the degree of predictability of the magnitude of one variable, given the magnitude of the other. In the problem of Rorschach validation, then, the usefulness of the results will depend upon the criterion chosen. The requisites for an acceptable criterion variable are: (1) amenability to the assignment of numerals with a high degree of reliability which represents a relative statement of the degree of presence of the variable, and (2) known or hypothesized relationships between the criterion variable and other variables useful in work with personality. Use of the term "numerals" in (1) above indicates another bias. The writer believes that the variables of personality are lawfully related to relevant environmental variables and since laws are numerical statements for optimal serviceability, it is necessary to quantify the variables in order to find the laws. This is deemed a requirement for a science of psychology. Finding the laws depends in part upon a methodological approach geared to finding lawful relationships. It then becomes an empirical matter as to whether or not the lawfulness is "out there."

In the case of the Rorschach it is held as a working hypothesis that many of the variables of personality can be measured by it, the task being to ascertain the tenability of the hypothesis by making a series of verifiable hypotheses relating clearly defined personality variables to each of the factors or groups of factors found in a Rorschach protocol. It is also held that the laws describing the functioning of the whole personality are of the same class as those relating component parts to environmental variables. In describing the whole personality, what is needed is a composite law expressing the component parts and their interaction. Again, this may be true or false according to future empirical findings. Its immediate value lies in focusing on a search for relationships that can be expressed in a single system of laws. As a practical matter, it is probably more fruitful to start with component



parts until more of the relevant variables for the total personality are ascertained. The rest of this paper will be devoted to an application of this orientation in an attempt to clarify some of the methodological problems inherent in Rorschach validation.

### *Levels of Inference*

From a similar theoretical and methodological background, Steisel and Cohen (34) have elaborated on Beck's differentiation of the levels of inference used in interpreting from Rorschach protocols. To clarify the problems of Rorschach validation they differentiate three levels of inference, each of which require validation.

Level I covers the inferences drawn from the raw response record for purposes of arriving at the appropriate scoring symbols and constructing an adequate psychogram.

Level II represents the descriptive statements which can be made by inspection of the psychogram, sequence analysis and analysis of specific verbalizations.

Level III includes the comprehensive, coherent personality diagnoses which may include the classification of the subject in some nosological group.

This can be considered a vertical schema ascending in order of degree of complexity (the number of variables involved in attempts to validate the inferences made). However, it can be pointed out that Level I does not involve the problem of validity as herein defined. Steisel and Cohen's Level I can more adequately be defined as entailing the problem of reliability, that is, the setting up of symbols each with a definition which summarizes certain aspects of a given response. The problem is to find the definition which will yield the highest degree of agreement in the case where several observers using the same definition for a given symbol are tested for the degree of concurrence of their applications of the symbol in question.

Having established that certain symbols can be applied with a satisfactory degree of reliability, Level II is reached. Level II represents the inferences as to specific processes or mechanisms which can be made about a personality. The raw data for these are found mostly in the psychograms (factors such as  $F + \%$ ,  $M$ ,  $C$ ,  $W : M$ , etc.) though material such as symbolism, pathognomic signs, and sequence analysis (19) may also be used. As a point of departure some of the Rorschach studies sometimes referred to as validation studies in the literature will be classified in terms of the Rorschach variables used and the method employed. Each study will be evaluated in terms of the notion of validity expressed above.

LEVEL I: PSYCHOLOGICAL PROCESSES  
REPRESENTED BY PSYCHOGRAM FACTORS

1. *Hypnosis*. Sarbin (31) took Rorschach protocols from a single subject under four different conditions. Two of these conditions were with the hypnotic suggestion that the patient was Mme. Curie, and then Mae West; a third condition was that of hypnosis with no suggestion; and the fourth condition was in the waking state. He found that each of the three artificial situations produced changes in several of the psychogram factors and he attributed the changes to induced sets. However, this does not qualify as a validity study in that the relationship between the particular stimuli used to induce the sets, and changes resulting, are not known.

Levine, Grassi, and Gerson (20) used hypnosis in a Rorschach study with an improved method. They obtained a Rorschach protocol under normal conditions and then under several conditions of hypnotically induced affective changes. They suggested certain situations (under hypnosis) the psychological consequences of which were predicted on the basis of clinical experience. The Rorschach protocols obtained under the experimental conditions reflected changes in psychogram factors for each of the induced conditions. Further, the changes obtained in each case reflected the Rorschach factors usually associated with the respective affective conditions induced. Validity of the results in this case, is, of course, dependent upon the accuracy with which a group of judges can predict the psychological correlates of certain situations. Accurate prediction in such cases is dependent in part upon knowledge of the differential reactions of a large group of subjects to the same situation. Thus the relationship between the factors of the emotionally charged situation and the personality involved must be known in order to predict for small groups. Also, since it is known that hypnotizability is related to certain aspects of personality, results of such studies cannot legitimately be generalized beyond groups that can be hypnotized.

Bergmann, Graham, and Leavitt (4) used Rorschach protocols in conjunction with hypnotic age level regressions. The authors do not claim this to be a Rorschach validity study. They acknowledge using Rorschach personality pictures obtained at the various suggested age levels to validate the extent of regression, in that superficial regression would not be expected to reflect a complete change in Rorschach patterns. However, were the extent of regression under hypnosis established by use of independent measures other than Rorschach, then Rorschach factors could be validated by this procedure. Detailed histories of certain levels of individual development could be obtained and a group selected on the basis of optimum uniformity.



On the basis of the personality features held in common for the individuals of this group, predictions could be made as to the relevant Rorschach factors that would be obtained under hypnotic regression to the respective levels of similarity. However, such a roundabout method is warranted only if there is a demonstrable need for an approach to the study of personality such as that used by Bergmann, Graham, and Leavitt.

In general, it might be said of use of hypnosis for Rorschach validation that considerably more data are needed to demonstrate how much of the personality changes are produced by the nature of the situation (i.e., effect of submission of the passive individual to the assertive hypnotist) and how much by the specific suggestions used.

2. *Drugs.* The studies reported in the literature (13, 18, 27, 37, 38) relating Rorschach factors to administration of drugs do not qualify as validation studies in that the laws describing the relationship between the drugs used (mescaline, histamine, amytal) and personality factors measured independently of the Rorschach are not known. The data reported suggest that the reactions obtained depend in part upon features of the predrug personality. Guttman (13) states, "... there is need of a searching investigation in terms of the subject's setting and individual equipment before specificity of reaction can be established." Until the laws relating a single drug with personality reactions are found, this method will not be serviceable in Rorschach validation work.

3. *Operationally Defined Situations.* As Williams points out (39) this method requires: (1) setting up of an "independent and standardized criterion situation which would yield a quantitative index to an operationally defined aspect of personality, and (2) examination of the validity of those Rorschach factors primarily associated with this personality process in terms of the criterion measure." He assembled the definitions offered for intellectual control and designed a situation to measure this. He used as his measure decrement in performance of a task requiring intellectual factors under stress compared with performance under optimum conditions. Thus, the definition of intellectual control is in terms of the operations used to measure it. He administered a standard Rorschach to each of his subjects and used as the Rorschach index of intellectual control  $F + \%$  and integration of form with color on the basis of hypotheses stated by Beck and Rapaport. He acknowledged the fact that other factors also were used in Rorschach interpretation and that he was operating at what is considered in this paper to be Level II. His results tend to bear out the hypothesized relationships between the selected Rorschach factors and the operationally defined aspects of personality. Structurally, this study conforms with the tenets presented here. However, he overlooked the fact that a linear rela-

tionship between intellectual control and  $F + \%$  is not hypothesized by Beck and Rapaport for emotionally charged situations. An  $F + \%$  above the optimum hypothetically represents an overcontrol due to inhibition and anxiety, compulsiveness, and rigidity or strong superego. Consequently, it might be hypothesized: (1) that overly-acute form perception is a defense against affective instability and as such might be related to loss of control under affective stress, and (2) that the threshold for affective stress would be lower than that for an individual with a lower but optimal  $F + \%$ . Therefore, it might be more in line with expectancy to adjust for  $F + \%$  above the empirical mean for a standardization population, rather than consider it to be a continuous distribution with respect to intellectual control under stress.

Another study with a similar methodological basis provides a focus for the problem of the usefulness of the independent variable with which the Rorschach factors are correlated. Steisel (33) attempted to demonstrate a relationship between the factors on the Rorschach commonly hypothesized to be related to suggestibility, and performance in the autokinetic and Hull body sway situations. The relationship between suggestibility and performance in these situations is not as universally acceptable as is the relationship between intellectual control and the situation used by Williams. Therefore, acceptance of Steisel's results as bearing upon suggestibility as a personality variable depends upon the demonstration of a relationship between his two criterion measures and another independent measure of suggestibility. Insofar as his concept of suggestibility is operationally defined in terms of these two indices it is a defensible method at the current stage of development of Rorschach validation; i.e., if he had found a relationship between his selected variables it would then be a fruitful problem to experiment with the autokinetic phenomenon and Hull body sway technique to find their relationship to other independently defined aspects of personality. Even if they should be found related to aspects other than suggestibility, Steisel's data would be useful.

Rockwell, Welch, Kubis, and Fisichelli (28) report an experiment concerned with color shock. They define shock in terms of the neurological conception as a condition of lowered excitability. They measured shock by taking a continuous reading of the palmar skin response during Rorschach administration and described the changes expected with "shock" present and those expected with startle. Thus, their measure constitutes an operational criterion for shock and startle. They found that shock, as neurophysiologically conceived, does occur on those cards and in those subjects where it is expected, whereas startle does not. On the basis of these results they reject the Beck criteria for "color shock" which are based on expectation of



a startle reaction to the color cards. This is a rather clear-cut example of an operationally defined situation devised to test two antithetical hypotheses. Acceptance of their interpretation depends upon evidence for relationship between the neurological conception of shock and Rorschach's conception of emotional inhibition. Assuming this relationship, it then remains to be determined what conditions the startle response (or more properly the Rorschach criteria for its presence) is related to—as a problem for those interpreters who use the “startle” criteria.

4. *Empirical Sign Studies.* The studies that fall into this group are concerned with demonstration of a relationship between certain single psychogram factors or groups of factors and effectiveness of various therapeutic techniques. (The sign studies concerned with relationship of Rorschach factors to psychiatric diagnostic categories will be dealt with in the series of Level III studies.)

In 1940 Piotrowski (23) published data which resulted from a study of two groups of schizophrenics, one of which showed improvement following insulin therapy while the other did not. He compared the pretreatment Rorschach protocols, and found that the presence of a response which included color as a determinant (*FC*, *CF*, or *C*) and for which there is a concrete or emotional association, i.e., other than color denomination, occurred significantly more frequently in the records of those who subsequently improved than those who did not. He reported that 81 per cent of the improved group met this criterion. Halpern (14) in a study of similar design found that the pretreatment Rorschach psychograms of those patients who subsequently improved with insulin therapy manifested greater productivity and those factors associated with “wider emotional range and greater capacity for empathy.” Graham (9) with a study similarly designed found that those who improve with insulin give more responses determined by the chiaroscuro aspects of the inkblots. Benjamin (3) gives a preliminary report of a follow-up study of insulin-treated schizophrenics for whom he had pretreatment Rorschach protocols. Though his data are unquantified, he states that the preliminary suggestion is that the appearance of low *F* + %, *DdW* and *DW*, and highly irregular sequence are associated with poor prognosis.

In these studies the criterion variable is judged improvement, and consequently judgment of improvement must be based upon clear-cut reliable criteria. Also, a clear statistical statement of the relationship between degrees of improvement and the signs reported to discriminate between “improved” and “unimproved” is required for an acceptable statement of validity. The above studies do not meet these criteria, but the results suggest possible relationships to be checked by more rigorous methods.

In an investigation concerned with factors related to success of metrazol

therapy with a group of psychotics, Morris (22) improved on the method in the preceding studies. He provided for more clear-cut criteria for judgment of improvement and then selected critical points for his signs by referring his findings to the chi-square distribution for a statistical statement of his results.

The significance of this group of studies lies in the attempt to establish the usefulness of the Rorschach for prediction under restricted, specified conditions. However, they contribute little to validation of the Rorschach in terms of the second validity criterion, which calls for correlation between Rorschach variables and independent measures of personality for which a relationship with other personality measures is known. Should it later be demonstrated that improvement following insulin or metrazol administration is related to certain dimensions of personality, then the signs found in the above studies can be used as a basis for statements about these new dimensions insofar as the studies from which they result are experimentally sound. It should be pointed out that though these studies may not be of immediate optimal value for Rorschach validation, this does not mean they are equally questionable from the standpoint of other theoretical and/or practical considerations, such as the ones for which they were designed. They are on the same continuum only with respect to the first criterion of validity and are cited here merely to clarify the issues involved in Rorschach validation per se.

## LEVEL II: PSYCHOLOGICAL PROCESSES REPRESENTED BY FACTORS OTHER THAN THOSE IN THE PSYCHOGRAM

1. *Symbolism.* Earl (7) reported an experiment in which he investigated the validity of the usual interpretations for certain content categories frequently obtained in the Rorschach protocols of disturbed subjects. He was primarily interested in water responses and selected a group of feeble-minded males aged thirteen to fifteen who were inmates of a residential school and had markedly immature personalities with severe conflict over masturbation. He selected those words with water content and all content words which occurred with the determinants *M*, *FM*, *m*, *K*, *FK*, and *KF* in the pre-experiment protocols. He then put his subjects in an hypnotic trance and told them, "I am going to tell you to think of something. As soon as you think of it something quite different will come into your mind. . . . I want you to say at once what it is. . . ." He then gave as stimulus words the words taken from the subject's own record which were suspected of being symbolic, interspersed among banal responses from his own record. He found that all



the crucial words used for expressing a response the determinant of which was *m*, had symbolic significance. All responses having water as primary or secondary content "produced anxiety symbolized or overt in the hypnotized subjects." In interpreting his study he suggests that no common symbolic significance exists, at least among children, and that these responses, however determined, are associated with profound anxiety and it may be suspected that primitive sexual stirrings are always present.

The nature of his sample and small *N* (5 subjects) places serious restrictions on interpretations of these data. It would be desirable to have a control group of "stable" individuals so as to measure relative occurrence of water responses, and to compare the associations offered in the experimental situations by those "normal" cases in which water responses do occur with the responses of the experimental group. It might also be possible to do this type of work by use of word association without hypnosis, thereby freeing it from the limitations of hypnotic validation studies discussed above. By use of such a method, it is reasonable to expect that affective disturbances associated with the responses suspected of symbolic usage could be demonstrated. However, it is a more difficult problem to ascertain consistency for the particular material being symbolized by a given symbol. Demonstration of such consistency may require an approach which integrates data for responses with data on behavioral indices which are associated with aspects of the inkblots. An example of an attempt to associate a behavioral index with features of the blot is Blake's empirical study of ocular activity (5) during administration of the Rorschach technique. He found that number and duration of fixations was related to properties of the cards. Card VI, for example, elicits the largest number of fixations for the top center detail. Card VI as a whole also elicits longer fixations during the initial inspection period than is true of the other cards in the sequence. It is also known that the top center detail of Card VI elicits the largest number of manifest sex responses and is suspected of producing symbolized sex responses as well as disturbances of sequence, form, control, etc., which are referred to as "sex shock." Therefore, it might prove fruitful to select those areas of the several cards which yield the highest number of fixations and use the responses offered for them in a method similar to that used by Earl. One might use a large group of normals and investigate the number of times a particular word is substituted in the word association experiment for the response offered during Rorschach administration.

2. *Empirical Sign Studies.* In 1941 Piotrowski (24) published a paper which presented six signs relevant to predicting the prognosis of schizophrenics given insulin therapy. It is unlike the similar studies reported above in that five of his signs are not found in the usual psychogram. He

presented six signs which are reported to be indicators of a good prognosis: Generic Term, Variety, Evidence, Color Response, Indirect Color Approach, and Demurring. He offers what seems to be adequate criteria for determining presence of responses with these attributes, designates the method of determining improvement (psychiatric staff decision), and claims 93.3 per cent correct prediction by use of three or more signs as the criterion for good prognosis. As in the other studies of this type it remains to be demonstrated that prognosis for schizophrenics with insulin therapy is lawfully related to certain specified personality variables or processes. On the basis of his own data, Piotrowski offers an hypothesis in terms of intellectual regression to account for the differences between schizophrenics who improve with insulin and those who do not. Those who improve have undergone emotional regression; those who do not improve have undergone intellectual regression during their pretreatment illness. Assuming that there are operations with which to measure intellectual regression (other than Rorschach variables), this hypothesis then becomes subject to experimental verification. If verified, Piotrowski's signs would then become valid indicators of emotional regression (though the converse could not be assumed *a priori* in the case of absence of the signs). This is an illustration of how a sign study which, though not of itself a complete validation process, might lead to a series of studies the end result of which is validation of Rorschach variables. However, before results such as Piotrowski's are followed up, it would be advisable to select a new group of schizophrenics about to be treated with insulin shock and test the accuracy of the prediction of his signs. This would guard somewhat against the possibility that his original results were influenced by variables other than those to which he attributed his results.

### LEVEL III: COMPREHENSIVE PERSONALITY DIAGNOSES

At this level an attempt is made to relate variables of the Rorschach protocol to groups with various diagnostic labels. The emphasis is on a complete personality pattern rather than the component processes as far as the validity criterion is concerned. Thus one might measure the degree of relationship between diagnosis arrived at by use of the Rorschach and diagnoses made by any other means such as other clinical tests, psychiatric observation, a combination of tests, interviews and observations, or case history. Several such studies have been reported in the literature with results supposedly supporting the validity of the Rorschach as a diagnostic tool. However, the instability of current diagnostic groupings makes them unsuitable for use as validity criteria. Diagnostic practices differ from institu-



tion to institution and even from psychiatrist to psychiatrist in the same institution. A further criticism of this type of study can be made in terms of efficiency which would hold even if stability of diagnostic groupings were granted. The number of variables involved in arriving at a Rorschach diagnosis is large. Some of them might be highly related to the criterion and some not at all related. If the relationship of each variable (used in the Rorschach diagnosis) to the criterion variable is not known, then no statements can be made on possible improvements of the method of diagnosis. It is an all or none conclusion where there is ample reason for suspecting that some of the variables in combination might be more effectively used than others. Guirdham (10, 11, 12) has indicated one way to handle this problem in his work with epileptics and depressives in which he sought complexes of Rorschach variables each of which had some power to differentiate the groups in question from other nosological groups. He empirically tried numerical relationships between these variables until he found an index which provided optimum differentiation. The logical development of such an approach would lead to statistical analysis of a large number of Rorschach records from a variety of diagnostic groups, with each Rorschach variable being put into a separate multiple regression equation for each diagnostic category. Such an approach would constitute a sign method and, although a high order one, would still be subject to the limitations of sign approach as a validity method discussed earlier in this paper (i.e., the variables used for prediction are not selected on the basis of degree to which they enter into relationships with other variables of personality).

Piotrowski (25) has outlined three approaches to Rorschach diagnosis of mild forms of schizophrenia which can be applied to the general problems of Level III validation. He discusses first pathognomonic signs such as contamination, fluctuations in form level, and position responses. The limitations of this approach he lists as use of too small a part of the total record, and the observed fact that a large porportion of schizophrenics produce none of the signs, and thus absence of the signs does not mean schizophrenia is not present. He takes up next tabular diagnostic procedures which are based on Rorschach components that can be counted. It consists in establishing the differences in the frequency with which the scoring symbols occur in the records of schizophrenic and nonschizophrenic subjects. He indicates his preference for what he calls a systematic diagnostic procedure which "rests on principles which interrelate the components of the Rorschach method, expressing their mutual dependencies, and arranging them into a dynamic system." By this means he implies that the interrelated variables he uses are based on valid interrelationships among Rorschach symbols, e.g.,

that the more human percepts, the larger the percentage of sharply perceived forms.

His criticism of the pathognomonic and tabular procedures are well taken as far as he goes, but he overlooks the fact that the interrelationships used for the systematic procedure are not validated in the strict sense. Their "validation" rests on unquantified clinical observations. Further, the search for such relationships as used by Piotrowski requires a deduction as to component processes from what is known about the unit personality of schizophrenics. Such a search would be warranted only if the component processes for a given diagnosis were known. With regularly occurring relationships between given diagnostic categories and component processes we would have stable diagnostic categories, provided there were also operations to measure them. If these operations were independent of the Rorschach variables. However, in such an ideal situation Rorschach validation would then conform with the description for Level validation.

### *Conclusions*

In view of the foregoing considerations, the position is taken here that Rorschach validating procedures can be most fruitfully treated as problems in relating Rorschach variables to independent measures of component personality processes. There is no body of facts that can be invoked to prove this contention. It rests on principles of methodology concerned with the demonstration of valid relationships which are acceptable within a scientific framework. The merit of this approach applied to the Rorschach will be substantiated or discredited by the results obtained from its use. If procedures are used with provision for precise statement of results which are repeatedly obtained under the specified conditions, the Rorschach becomes a surer clinical device and it will facilitate psychological research. The clinician frequently protests that the experimental psychologist deals with artificial situations which are in no way related to the behavior which must be dealt with in the clinical problems of diagnosis and treatment. If the relationship between one or more Rorschach variables and component personality factors could be demonstrated, then those factors could be subjected to controlled experimentation with the Rorschach used to measure the changes that occur under varied conditions. It might also be possible to develop relationships between certain manipulable conditions and the origin of some of the personality components. Thus a single measuring instrument, satisfactorily validated, could help to establish a connection between experimental results and clinical problems. It would seem particularly useful to



aim toward this possibility in view of the widely held hypothesis that the Rorschach test samples a larger number of variables of personality than any other single instrument.

In conclusion, it should be made explicit that the point of view on validation presented here does not do violence to the experienced interpreter's demands that each Rorschach factor be interpreted in terms of the other factors for any given record. What is required is specific statement of any particular constellation that is believed to be related to some measurable aspect of behavior, and then an appropriate test of this statement.

## BIBLIOGRAPHY

1. BECK, S. J. Problems of further research in the Rorschach test. *Amer. J. Orthopsychiat.*, 1935, 5, 100-115.
2. BECK, S. J. Error, symbol and method in the Rorschach test. *J. Abnorm. Soc. Psychol.*, 1942, 37, 83-103.
3. BENJAMIN, J. D. A method for distinguishing and evaluating formal thinking disorders in schizophrenia. In J. S. Kasanin (Ed.), *Language and thought in schizophrenia*. Berkeley and Los Angeles: Univ. of California Press, 1946.
4. BERGMANN, M. S., GRAHAM, H., & LEAVITT, H. C. Rorschach exploration of consecutive hypnotic age level regressions. *Psychosom. Med.*, 1947, 9, 20-28.
5. BLAKE, R. R. Ocular activity during administration of the Rorschach test. *J. Clin. Psychol.*, 1948, 4, 159-170.
6. BROSN, H. W., & FROMM, ERIKA O. Some principles of Gestalt psychology in the Rorschach experiment. *Rorschach Res. Exch.*, 1942, 6, 1-15.
7. EARL, C. J. C. A note on the validity of certain Rorschach symbols. *Rorschach Res. Exch.*, 1941, 5, 51-61.
8. FRANK, L. K. Projective methods for the study of personality. *J. Psychol.*, 1939, 8, 389-413.
9. GRAHAM, VIRGINIA T. Psychological studies of hypoglycemia therapy. *J. Psychol.*, 1940, 10, 327-358.
10. GUIRDHAM, A. The Rorschach test in epileptics. *J. Ment. Sci.*, 1955, 81, 870-893.
11. GUIRDHAM, A. Simple psychological data in melancholia. *J. Ment. Sci.*, 1936, 82, 649-653.
12. GUIRDHAM, A. The diagnosis of depression by the Rorschach test. *Brit. J. Med. Psychol.*, 1936, 16, 130-145.
13. GUTTMAN, E. Artificial psychoses produced by mescaline. *J. Ment. Sci.*, 1936, 82, 203-221.
14. HALPERN, FLORENCE. Rorschach interpretation of the personality structure of schizophrenics who benefit from insulin therapy. *Psychiat. Quart.*, 1940, 14, 826-833.

15. HARRISON, R. The thematic apperception and Rorschach methods of personality investigation in clinical practice. *J. Psychol.*, 1943, 15, 49-74.
16. HERTZ, MARGUERITE R. Rorschach: twenty years after. *Psychol. Bull.*, 1942, 49, 529-572.
17. HERTZ, MARGUERITE R. The Rorschach method: science or mystery. *J. Consult. Psychol.*, 1943, 7, 67-80.
18. KELLEY, D. M., LEVINE, K., PEMBERTON, W., & LILLIAN, K. K. Intravenous sodium amytal medication as an aid to the Rorschach method. *Psychiat. Quart.*, 1941, 15, 68-73.
19. KLOPPER, B., & KELLEY, D. *The Rorschach Technique*. New York: World Book Co., 1942.
20. LEVINE, K. N., GRASSI, J. R., & GERSON, M. J. Hypnotically induced mood changes in the verbal and graphic Rorschach: a case study. *Rorschach Res. Exch.*, 1943, 7, 130-144.
21. MACFARLANE, JEAN W. Problems of validation inherent in projective methods. *Amer. J. Orthopsychiat.*, 1942, 12, 405-410.
22. MORRIS, W. W. Prognostic possibilities of the Rorschach method in metrazol therapy. *Amer. J. Psychiat.*, 1943, 100, 222-230.
23. PIOTROWSKI, Z. A. A simple experimental device for the prediction of outcome of insulin therapy in schizophrenia. *Psychiat. Quart.*, 1940, 14, 267-273.
24. PIOTROWSKI, Z. A. The Rorschach method as a prognostic aid in the insulin shock treatment of schizophrenics. *Psychiat. Quart.*, 1941, 15, 807-822.
25. PIOTROWSKI, Z. A. Experimental psychological diagnosis of mild forms of schizophrenia. *Rorschach Res. Exch.*, 1945, 9, 189-200.
26. RAPAPORT, D. In Rorschach forum report. *Rorschach Res. Exch.*, 1939, 3, 107-110.
27. ROBB, R. W., KOVITZ, B., & RAPAPORT, D. Histamine in the treatment of psychosis. *Amer. J. Psychiat.*, 1940, 97, 601-610.
28. ROCKWELL, P. V., WELCH, L., KUBIS, J., & FISICHELLI, V. Changes in palmar skin resistance during the Rorschach test. I. Color shock and psychoneurotic reactions. *Monthly Rev. Psychiat. Neurol.*, 1947, 113, 9-152.
29. RORSCHACH, H. *Psychodiagnostics*. (Trans. by P. Lemkau & B. Kronenberg.) New York: Grune and Stratton (distr.), 1942.
30. ROSENZWEIG, S. Outline of a cooperative project for validating the Rorschach test. *Amer. J. Orthopsychiat.*, 1935, 5, 121-123.
31. SARBIN, T. R. Rorschach patterns under hypnosis. *Amer. J. Orthopsychiat.*, 1939, 9, 315-318.
32. SARGENT, HELEN D., Projective methods: their origins, theory and applications in personality research. *Psychol. Bull.*, 1945, 42, 257-293.
33. STEISEL, I. An experimental investigation of the relationships between some measures of the Rorschach test and certain measures of suggestibility. Unpublished doctoral dissertation, Univ. Iowa, 1949.
34. STEISEL, I., & COHEN, B. D. The problem of validation of the Rorschach test with special reference to the method of direct experimentation. Unpublished paper, 1947.



35. THURSTONE, L. L., The Rorschach in psychological science. *J. Abnorm. Soc. Psychol.*, 1948, 43, 471-475.
36. VARVEL, W. A. Suggestions toward the experimental validation of the Rorschach test. *Bull. Menninger Clin.*, 1937, 1, 220-226.
37. WERTHAM, F., & BLEULER, M. Inconstancy of the formal structure of the personality; experimental study of the influence of mescaline on the Rorschach test. *Arch. Neurol. Psychiat.*, 1932, 28, 52-70.
38. WILKINS, W. L., & ADAMS, A. J. The use of the Rorschach test under sodium amyltal and under hypnosis in military psychiatry. *J. Gen. Psychol.*, 1947, 36, 131-138.
39. WILLIAMS, M. An experimental study of intellectual control under stress and associated factors. *J. Consult. Psychol.*, 1947, 11, 21-29.

James O. Palmer

# A DUAL APPROACH TO RORSCHACH VALIDATION

## *A Methodological Study*

### *I. The Two Approaches— A General Statement of the Problem<sup>1</sup>*

#### A. INTRODUCTION

THE NUMEROUS investigations of the validity of the Rorschach have been reviewed very thoroughly in three articles by Hertz (5, 6, 7), and the validity of the TAT has been similarly summarized by Tomkins (19). However, a brief restatement of the methods used by various investigators in establishing the validity of projective techniques may serve as an orientation to the particular questions considered in the present study. Thus far, the authorities who

Reprinted from *Psychol. Monogr.*, 1951, 8 (Whole No. 325), 1-27, by permission of the American Psychological Association and the author.

1. This study was conducted under the auspices of the Veterans Administration Regional Office, San Francisco, while the author was in training in the Clinical Psychology Training Program of the Veterans Administration. The author is therefore deeply indebted to the Veterans Administration for making this study possible. The opinions stated in this report are, however, those of the author and do not necessarily reflect the viewpoint of the Veterans Administration.



have reviewed this problem have been concerned primarily with the types of evidence used for the validation of projective techniques. Hertz (5) distinguished four main types of validation studies, as providing different kinds of evidence: (a) *clinical studies*, in which the usefulness of these techniques is illustrated in the analysis of case histories, therapy, etc.; (b) *experimental studies*, in which changes in the test results are shown to accompany controlled changes in the individual's pattern of behavior; (c) *studies of defined groups*, in which certain patterns of test results are established as associated with the characteristic behavior of known groups; and (d) *predictive studies*, in which the degree of agreement is measured between the description of personality derived from the results of a projective technique and that obtained from an analysis of some criterion, for example, a life history.

Although various writers (Tomkins [17], Macfarlane [12], and Symonds and Krugman [16]) have granted the possibility of approaching the validation of projective techniques in various ways, they have, at the same time, been careful to emphasize that the chosen method must take into account the nature of the technique being validated, particularly the concept of personality underlying the use of this technique. The most comprehensive argument concerning this point has been presented by Frank (3) in his classic discussion of the scientific basis of projective techniques. He pointed out that the "personality" which projective techniques are designed to evaluate is a *framework of* intervening concepts, a framework that relates the details of the individual's manifest behavior in terms of a *pattern of* motivations and attitudes. Macfarlane (12) also has considered this use of *interrelated* constructs to be a central problem "inherent in the validation of projective techniques." The point stressed by Frank is that personality as a configuration of functioning processes cannot be meaningfully broken up into isolated traits, but that part functions can only be described in terms of their interrelationships within the whole pattern. From this viewpoint, a description of personality derived from the results of a projective technique would require a method of validation which could test the accuracy of this description as a whole unit.

This assumption concerning the relationship between personality descriptions derived from projective techniques, and the method employed for their validation constituted the point of departure of the present research. In the light of this concept of personality, two divergent predictive approaches to validation were applied to an established projective technique, the Rorschach. One of these approaches, the matching method, was designed specifically to test the validity of description of personality as whole units. The other approach, which attempts to validate these descriptions

item by item, does not necessarily take into account the Gestalt nature of these descriptions. The general intent of this study was to test the relative applicability of these two methods in investigating the validity of a projective technique.

## B. THE INTERPRETATION AS THE OBJECT OF VALIDATION

Before proceeding with the description of these two methods, it may be well to emphasize that this study is concerned with methods of validating the *interpretation* or description of personality as derived from the responses of the subject, rather than with consideration of specific *scores* or discrete responses. While there are merits in dealing with the so-called objective data of projective tests, this author agrees with such authorities as Hertz (6) and Marfarlane (12) that behind such scoring systems lie implicit assumptions about personality functioning. It thus appears to this author more reasonable to deal directly with these interpretative assumptions and avoid the current controversies concerning scoring categories and their discrete meanings. The question of whether or not a set of idiosyncratic responses represents in a rough manner the general pattern of functioning of the individual, and of how this question may be answered, seems a legitimate object of study.

## C. THE MATCHING APPROACH

It was clearly apparent to Vernon (19), even during the developmental stage of projective techniques, that there was a need for a statistical approach which would treat their interpretation as a single, whole unit. With this specific problem in mind, he developed what has become known in the literature as the *matching method*. As Vernon (18), Hunter (9), and Krugman (11) have used it, this method consists of the following procedures: An interpretative report from a projective technique and a case analysis are prepared, independently, for each individual in a given sample. The sample is divided into small groups, ranging from five to ten subjects, known as *matching groups*. The interpretative reports and case analyses of each group<sup>2</sup> are presented, unidentified as to subject, to several judges who then attempt to match each of the test reports to the corresponding case analysis of the same individual's life history. Validity is then expressed in terms of the success of this matching. Chapman (1) derived the statistics for deter-

2. While most investigators have used matching groups consisting of an equal number of reports and case analyses, the matching groups may be uneven, e.g., ten interpretations to five analyses, or five to one.



mining the chance variation and the significance of the success of matching. Vernon (19) has added a formula for a coefficient of contingency,  $C$ , permitting a statement of the degree of relationship between the test and the criterion as implied in the success of the matching.

Vernon (18) admitted that his method was "only a coarse beginning" to the validation of projective techniques and suggested two additional steps to this procedure: (a) the homogeneity of the matching groups must be determined (obviously, a group of very similar reports would be more difficult to match than a group of very dissimilar reports); and (b) the reliability of the matching judges should be determined.

The application of the matching method to the validation of projective techniques has produced varying results. In his original study on the Rorschach, Vernon (18) reported an average contingency coefficient,  $C$ , of  $.833 \pm .0315$ . Vernon noted that "the actual size of the  $C$  depends very largely on the degree of heterogeneity or distinctiveness of the subjects in each group. As far as possible, a normal degree of heterogeneity was aimed at" by randomly selecting the cases for the groups out of the whole sample (18, p. 213). However, Vernon's matching groups may have been more heterogeneous than he assumed, as is suggested by the results of later studies, Hunter (9), in a study of fifty school children, reported that only five Rorschach reports were matched correctly by all four judges, and that each judge, singly, was successful in only 30 to 40 per cent of the matchings. She concluded, therefore, that matching was "of doubtful value" . . . "Calculated to differentiate only extreme cases" (9). Although this investigator did not state her method of selecting the matching groups, she did remark that the personality sketches were all very similar.

The chief objection to the matching method, however, is that it permits, at best, only a very general statement about the accuracy of an interpretative report, namely, the statement that on the whole the report is similar to the personality pattern depicted by the criterion. While it is reassuring to know that an essentially meaningful interpretation may be drawn from a projective technique, it would be even more satisfactory to know how well the personality configuration is delineated in an interpretative report. If the interpretative report is, in general, similar to the case analysis, successful matching may occur, even though the accuracy of many of the statements within the report is dubious. In fact, Cronbach (2) considered that the matching method depends too much on the presence or absence of small clues. A proponent of the matching method might argue that, if the same personality pattern is, in general, described in both the interpretative report and in the case analysis, then the statements about particular functions of the personality would be likely to be similar in both descriptions.

Unfortunately, the matching method does not provide any tests of this argument.

#### D. AN ITEM ANALYSIS METHOD

In a study of the validity of the TAT, Harrison (4) introduced a procedure which simultaneously checked both the *degree* and *area* of the accuracy of his interpretations. His interpreters wrote out an "itemized analysis" of the test protocols, i.e., lists of separate interpretative statements, and the case analyses were prepared in a similar fashion. The judges then compared the two sets of statements, *item by item*, denoting each item of the interpretation as "right," "wrong," or "?." This index of validity was significantly higher for his sample of interpretations than for a random group of interpretations, or for a group of "mock" reports, matched randomly with the same criterion. Cronbach (2)<sup>3</sup> described a validation design for projective techniques which is quite comparable to that employed by Harrison. Cronbach's conclusions, which might also be said to apply to Harrison's method, were that his type of approach (a) yielded a statistically sound test of significance, (b) "identifies objectively the accurate and inaccurate aspects of the prediction," and (c) permitted "identification of the types of cases for whom prediction is relatively accurate" (2, p. 373).

It should be noted that in this procedure, the proof of validity hinges on the premise that the judges accept the statements of the interpretation as being similar to the items in the case analysis. The criterion for being "similar" is not stated in either of these articles, and from this procedure, no conclusion can be drawn as to the *degree* of similarity between the two sets of items. This degree of similarity might be measured by a rating scale of agreement, as used by Krugman (11). The ultimate step in this validation procedure would be to demonstrate that the personality of the individual as inferred from the test and from a life situation could be described by the same statements, e.g., on a rating scale or on a check list of commonly used statements.

The feature of this "item analysis" method of greatest import to our discussion is that on the surface it contradicts the assumption by Frank (3) and Vernon (18, 19), namely, that, since these descriptions deal with an integrated personality structure, they could not be validated piece by piece. This seeming contradiction might be explained, however, by the hypothesis that *the validity of these separate items depends on the validity of the whole description*. In the strictest sense, this hypothesis would mean that separate statements within the interpretative report would be valid *only* if the whole

3. Since Cronbach's article was published after the present study was completed, the particular design which he introduced was not originally considered in this investigation.



interpretation were valid. At least, it would indicate that if the interpretation as a whole is accurate, then the isolated statements drawn from the interpretative reports would be *more likely* to be accurate. It should be emphasized that this hypothesis refers to the validation of interpretations which stress the *relationships* between the various functionings within the personality as a whole.

However, the studies of Harrison (4) and by Cronbach (2) do not supply any direct evidence in support of this hypothesis, since neither study provided a test of the accuracy of the interpretations as integrated, whole reports. Harrison did not state whether or not certain of his cases had significantly lower "validity indices." Although Cronbach concluded that his procedure allowed "the identification of the types of cases for whom prediction is relatively accurate," he did not describe these cases in the part of his study reported to date. It is possible that in both of these studies, some interpretations were inaccurate in the description of the personality as a whole, and that, therefore, the items drawn from these interpretations failed to attain significance.

Unfortunately, it cannot be determined from either Harrison's or Cronbach's articles exactly on what basis the interpretations were broken down into isolated items. One main criticism of these two studies is that no rationale is presented as a basis for the selection of the items. In fact, Harrison did not adopt a dynamic, structural approach to personality in his interpretations, but stated that his approach to personality was "eclectic and emphasized common sense psychology" in contrast to Murray's (13) theories or to psychoanalysis. Nor did Cronbach describe the theoretical bias of his Rorschach interpreters, although he did indicate that a more complete report would follow his introductory article. Exactly what types of statements about personality were validated, or might be validated, in this fashion remains undetermined. There is no assurance that this methodological design is applicable to the validation of interpretations based on a dynamic theory of personality.

#### E. PURPOSES AND PROCEDURES OF THE PRESENT STUDY

The purposes of the present study were:

1. To test the hypothesis (discussed above) that the validity of separate statements about personality, inferred from projective techniques, depends on the accuracy of the interpretation as a whole.
2. To determine whether a test of the validity of isolated statements is

applicable to interpretations based on a dynamic, structural concept of personality.

3. To determine whether the personality of the individual as inferred from the test protocol and from the criterion situation could be described by the same set of statements.

## *II. The Test, the Criterion, and the Sample*

### A. THE SELECTION OF THE RORSCHACH TECHNIQUE

In order to study the applicability of two methods of validation, it was essential that these approaches be tested on a projective technique of relatively accepted validity. After due consideration, the Rorschach technique was chosen, mainly on the strength of a comparatively longer and more varied background of validation studies. For a complete review of these investigations of the validity of the Rorschach, the reader is referred to the three articles by Hertz (5, 6, 7).

Predictive studies of the accuracy of Rorschach interpretation have not been as numerous, nor have the results been as uniformly positive, as those reported for the other types of approaches. Other than the studies using the matching method—which is under consideration here—most statistical studies have attempted to correlate isolated Rorschach signs with manifest behavior, usually with negative results. In regard to these studies, Hertz remarked, "The abortive dissection of the psychogram in the search for static factors in isolation has distorted the [Rorschach] method" (6, p. 549).

In addition to summarizing the various studies containing evidence of the validity of the Rorschach, Hertz presented many positive criticisms of these studies and recommendations for further investigation. In particular, she stressed the need for more experimental and differential studies to evaluate the various hypotheses underlying the interpretation of the Rorschach. The main purpose of her review was to stimulate sound experimental design in these studies.

1. *The Administration of the Test.* The method of administration followed the procedure described by Klopfer and Kelley (10). All the examiners took particular pains to secure a thorough "inquiry" into the features of the blots which elicited the subjects' responses. Probing and suggestive questions were avoided, however, until the "testing of the limits." Twenty-one of the subjects were tested by the author; the other seven subjects had been previously tested by other experienced administrators.

2. *The Protocols.* The validity of a Rorschach interpretation depends both on the quantity and quality of the subject's responses. Most of these



records offer a wealth of variegated responses as raw data for the interpreter. The completeness of the records was, to some extent, assured by the technique of administration, i.e., by the thorough inquiry. The quality of the protocols may also have been affected by the nature of the sample; possibly the patients had been selected for psychotherapy because they were comparatively more responsive and less restricted in their functioning.

## B. THE SUBJECTS

In three Veterans Administration installations, the Rorschach was administered to all patients who were currently receiving psychotherapy and to whom the Rorschach had not previously been given. Of these twenty-eight subjects, eleven were in a neuropsychiatric hospital, eleven were from an outpatient clinic, and six were attending a nearby university clinic. These subjects ranged in age from nineteen to forty-two years, with a mean age of twenty-eight years. Ten of the patients were classified as psychotic, sixteen as neurotic, and two had other diagnoses. Except for the differences in diagnostic classification (see further discussion below), the character of this sample did not appear to have any direct bearing on the study of these two methods of validation.

## C. THE CRITERION

Obviously, the validation of a projective technique depends on the use of a criterion description which is comparable in nature to the test interpretation, and which is based on an adequate study of the individual. While the functioning of the individual's personality may be inferred from his manifest behavior as summarized in a life history or factual case study, this functioning may be observed even more directly and intimately in a psychotherapeutic study, i.e., in the individual's expression of his feelings and attitudes during psychotherapy, and in his emotional reactions to the psychotherapeutic situation. This criterion has been used in at least two major clinical validations of projective techniques: Hertz and Rubenstein (8), and Tomkins (17). It was also recommended by Rosenzweig (15) in his outline for a comprehensive study of the validity of the Rorschach.

In the present study, the Rorschach interpretations were compared, in the two validation methods, with the therapists' impressions of their patients. The seventeen therapists acted as the judges in both validation experiments, i.e., they selected the Rorschach reports which matched their patients in the matching experiment and described their patients in terms of the choices on the item check list. Thus, the validation judges were able to evaluate the Rorschach interpretations on the basis of an intimate and

extensive knowledge of the subject, instead of having to reply on the basis of a summarized sketch compiled by a disinterested party.

The therapy which the patients were receiving was psychoanalytic in nature, i.e., its purpose was to reveal to the patient his unconscious attitudes and motivations through an analysis of his emotional reaction to the therapeutic situation itself. The purpose of this therapeutic study of the patient's underlying attitudes and motivations may be considered equivalent to the aim of Rorschach interpretation. In fact, these therapists frequently requested a Rorschach report on their patients' personalities as an aid in planning treatment (excepting the patients who were subjects of this experiment). A majority of the therapists were also expert in the administration and interpretation of the Rorschach technique.

In the main, the therapists' impressions of their patients were derived from frequent contact with them. The total number of therapeutic interviews at the time the therapist made his judgment ranged from 6 to 90, with a median of 19 interviews: only 4 subjects were interviewed less than 10 times, while 8 had been interviewed over 30 times. In 20 of the 28 cases, the Rorschach was administered when therapy was in a beginning stage, i.e., before the fifth interview: the median number of interviews at the time of testing was 1, with a range of 0 to 40. A median of 15 weeks elapsed between the time of testing and the time of judgment; in no case was there less than a 7-week interval, and in 2 cases, the interval was over 30 weeks. During this period, the therapists interviewed their patients between 5 and 80 times, with a median of 11 interviews occurring between testing and judging. As to the frequency of these interviews, 6 of the cases were seen 3 or more times weekly, 5 others were seen twice weekly, and all but 2 patients were seen regularly at least once a week. These 2 cases were interviewed frequently, but at irregular intervals. The minimum opportunity which a therapist had to observe his patient was 6 interviews, occurring at irregular intervals, over a period of 16 weeks. The maximum observation occurred in a case where the patient was interviewed 90 times, 3 times a week, over a 30-week period. In summary, it may be said that the therapists' impressions were derived after extensive and frequent contacts with the subjects and may be considered comparable, in their theoretical approach to personality, to the Rorschach interpretations.

### *III. The Matching Approach*

In brief, the matching experiment consisted of the following procedures:

1. Interpretative reports were prepared from each of the Rorschach records by one interpreter (the author).



2. The reliability of these reports was checked in two ways: by having the reports matched to the protocols, and by having them matched to a duplicate set of reports which had been prepared by another psychologist.

3. In order to test the reliability of the therapists in their matching technique, they were given a group of five sample interpretations from which they had to select the one which matched a corresponding sample case analysis.

4. For each patient, a matching group was chosen, consisting of the interpretation of that patient's Rorschach, referred to hereafter as the *experimental* interpretation; and of four other interpretative reports, to be referred to as *alternates*.

5. Each therapist was then asked to select, from the group of five reports, the one report which he believed most closely represented his patient; subsequently, the therapist made a second choice among the remaining four reports.

#### A. THE NATURE OF THE INTERPRETATIONS

In order to standardize the method of interpretation and the style of the interpretative reports, all the protocols were interpreted by one person, the author. These interpretations were based solely on an individual's responses to the test material and on his behavior during the administration of the test.<sup>4</sup> The method of interpretation followed that outlined in Klopfer and Kelley (10), particularly in the scoring of the responses and in the preliminary analysis of the psychogram. The conceptual framework employed throughout this process of interpretation was, broadly speaking, psychoanalytical. As far as possible, these descriptions were couched in everyday idiom, and both Rorschach and psychoanalytic terms were avoided.

#### B. THE RELIABILITY OF THE INTERPRETATIONS

Since all of the interpretative reports employed in the matching experiment were prepared by one person, it was necessary to determine whether these interpretations were reliable in the same sense that the consistency and accuracy with which one scores a psychometric instrument might be checked. As a rough test of this reliability, three judges, skilled in Rorschach interpretation, attempted to match the report to the protocols, in groups of five each. Since this study was directed at the reliability or consistency of fully verbalized interpretations rather than of standardized symbols or scores, no attempt was made to check the author's scoring. These judges were

4. The patient's behavior during the testing (as distinguished from his responses to the materials) was not recorded. Since most of the records were interpreted some time after the test administration, little if any account was taken of this factor.

instead presented with the *unscored* responses and asked to match these directly to the author's statements about the various subjects' personalities. All three judges were 100 per cent successful in this matching.

Despite this positive result, it was possible to question the reliability of these interpretations, i.e., whether they were similar to descriptions derived by some other interpreter. As a further check on this reliability, the protocols were interpreted independently by another psychologist.<sup>5</sup> Three judges matched, with complete success, the first five interpretations of this second set of reports with the five corresponding interpretations by the author. Since this result coincided with the results of Krugman's (11) more comprehensive study of Rorschach reliability, the success of this single matching was considered sufficient indication of the reliability of the interpretations used in the present study.

### C. THE RELIABILITY OF THE THERAPISTS IN MATCHING

Prior to the matching of the Rorschach interpretations in the main part of this research, the therapists were briefly trained in the use of the matching technique. In order to check the reliability of the therapists in matching, the author prepared a case analysis on a patient not included in the validation study. This patient's Rorschach record was interpreted independently by the interpreter who had participated in the study of the reliability of the interpretations. Using this case analysis as a criterion, ten of the therapists attempted to select this experimental interpretation from among four alternative interpretations (previously prepared by this other interpreter).

On this trial, six of the therapists matched the sample interpretation correctly on their first choice, and three others indicated it as their second choice. In this instance, successful matching on first choice could be expected by chance in two cases, i.e., two out of ten times. When first and second choices were considered, chance matchings might occur in four instances in ten matchings. According to the tables of "General Term of Poisson's Exponential Expansion" in Pearson (14)<sup>6</sup> the obtained results of both the

5. The author wishes to acknowledge the patient assistance of Mr. Mervin Freedman, Mr. Patrick Sullivan, and Mr. William Cook of The University of California who acted as the judges here, and of Mr. Allen Dittmann, who prepared the second set of reports.

6. Throughout this study, many of the resultant proportions of chance agreement were so small that their distribution was thought to be considerably skewed and platykurtic. The use of a standard error of a proportion and its interpretation in terms of the normal probability integral would have yielded erroneous probabilities. It was thought, however, that the computation of exact binomial probabilities would have required more effort than their usefulness justified, so approximations to these probabilities were obtained from the Poisson distributions. This distribution is useful in approximately binomial probabilities when  $p$  is small in comparison to  $q$ , but where the possible number (of agreements, in our case) is finite.



first choices alone (six correct matchings) and of the two choices combined (nine correct matchings) are significant beyond the 5 per cent level of confidence. Admittedly, this limited study of the reliability of the therapists in matching a single case was not completely comparable to the matching study of the validity of the twenty-eight cases, as described below. Still, in view of the positive results of this brief reliability study, it seems reasonable to expect that these therapists would be approximately reliable in other matching experiments—particularly one in which they would be more familiar with the criterion, i.e., their own patients.

#### D. SELECTION OF THE MATCHING GROUPS

As noted in Section I, the results of a matching study depends to a large extent on the variability among the descriptive reports which constitute the matching groups. In contrast to previous investigations which also made use of the matching method, the present research included an attempt to control the heterogeneity of the matching groups. This particular step in the present research may, therefore, bear some detailed explanation.

The purpose of this step in the matching procedure was to select matching groups having the same degree of heterogeneity. It seemed desirable that the *alternate interpretations* to be included in a group with the experimental interpretation *should be neither very different from nor very similar to that experimental interpretation*. To achieve this degree of heterogeneity, it was necessary to compare each interpretation with all other interpretations which might appear within a group as an alternate. In order to estimate the differences between interpretations, the following rough rating scale was adopted:

SS—Both interpretative reports describe the same basic personality features but may differ in specific characteristics.

S—Both reports describe similar basic personality features but may differ in specific characteristics.

SO—Both reports describe some similar basic features *and* some similar specific characteristics, but also differ slightly in both respects.

O—Both reports differ as to basic features but present some similar specific characteristics.

OO—Both reports differ in all respects.

N—The reports are not comparable.

For the sake of convenience, the first fifteen reports collected from the sample were rated on this scale prior to the last thirteen reports. Three judges<sup>7</sup> rated these reports on the above scale; each judge made an inde-

7. The author appreciates the assistance of Mr. Timothy Leary and Mr. Walter Klopfer in this task.

pendent rating first, followed by a final pooled rating by all three judges. These ratings were made on the over-all description of the personality rather than any specific cues. Thus, two interpretations discussing latent homosexual trends as important to the personality picture but differing in most other respects, i.e., in basic personality structures, might be rated at least *O*, if not *OO*.

In selecting the matching groups, reports which rated *SO* with the experimental report were given preference; a few reports compared as *O* or *S* were also used in some groups, but none of those extremely different or similar were included. Thus, the matching groups were of appropriate heterogeneity with regard to the experimental report.

#### E. THE SEQUENCE IN WHICH THE TWO APPROACHES WERE USED

Since both the matching and the check list judgments were made by the same judges (i.e., the therapists), the sequence in which the two techniques were tested carried a possible contamination: a therapist who matched his case before using the check list might be influenced by the selected report when the time came for him to make choices on the check list. Or, if he used the check list first, he might acquire a set for the matching of the report. Although such bias could not be wholly prevented, its possible effect was taken into account by systematically varying the order in which the therapist performed the two tasks. Therefore, in fourteen of the cases (seven in each half of the sample), the therapist selected the report before he made choices on the check list—the sequence being reversed in the other fourteen cases. The possibility of this type of bias was further lessened by the fact that the therapists always made the two judgments separately, with an intervening period of two to three weeks.

#### F. HOW THE THERAPIST MADE HIS SELECTIONS

Each therapist was presented with one matching group for each of his patients. Each group consisted of five reports (one of which was derived from his own patient's Rorschach). The therapist was instructed to select the one report which matched his patient. After this first choice was made, the therapist was asked to name a second choice from the remaining four reports. The selection of a second choice was requested in order to allow for partial errors in matching, particularly in those instances when a judge might be undecided as to which of two similar reports to select.

Another factor which had to be considered in this procedure was the



possibility that the patient's personality had been altered by the therapy which intervened between the time the Rorschach was administered and the time the therapist made his selection. Therefore, as he made his selection, the therapist was reminded of the date of the test administration, and he was asked to consider the patient's personality as it had been at that previous time.

#### G. RESULTS OF THE MATCHING

In eleven of the twenty-eight cases, the therapists correctly selected the interpretation of their patient's Rorschach as a first choice from the matching groups; only two more were correctly selected as second choices. In terms of chance expectancy (using Poisson's tables) this result is significant beyond the 3 per cent level of confidence. Using Vernon's (19) formula for the coefficient of contingency,  $C$  is equal to  $.434 \pm .078$ .

Although this matching was above chance, the relationship between the Rorschach and criterion, indicated by this  $C$ , was considerably lower than reported in previous studies. Vernon (18) found a  $C$  of  $.833 \pm .047$ .<sup>s</sup> Krugman (11) reported a  $C$  of .850. Both studies differ from the present investigation in two important aspects:

1. They did not specifically control the heterogeneity of their matching groups. It seems reasonable to presume that in the present study the control of this factor created a more difficult task for the judges, and consequently a more acute test of the Rorschach reports.

2. The previous studies used equal numbers of reports and case analyses, while the present investigation employed a five-to-one matching. Thus, in the present study, the judges were forced to differentiate among five reports, with only one criterion as a basis of judgment; in this sense, the chance of success was probably much smaller than in the previous studies.

Considering the factor of a more differentiating task for the judges, who therefore had less chance for successful matching, the degree of validity obtained in the present study may perhaps be regarded as comparatively more significant than the findings in the previous studies which did not include these controls.

The results of the matching experiment might also have been affected by other variables in the nature of the sample or in any of the procedures used in collecting and presenting the data. Seven variables were considered as possibly affecting the results of this matching, namely: (a) the type of installation (hospital or outpatient); (b) the psychiatric diagnosis (psychotic or "other neuropsychiatric disorder"); (c) the total numbers of interviews

8. Vernon reported a  $PE$  of .0314, converted here for purposes of comparison to a standard error.

(above or below the median of 19); (d) the numbers of interviews after testing (median or 11 interviews); (e) the frequency of the interviews (weekly or more frequent); (f) the order in which the Rorschachs were administered (a difference in results was indicated between the two halves of the sample); and (g) the judgment which the therapist made first (matching or check list). A study of the effect of these seven variables was made to discover if they had any possible relationship to the matching results (see Table 1).

TABLE 1  
Differences in the Proportions of Cases Matched Correctly, Between Various Characteristics of the Sample, and Between Various Procedures in Matching

Groups Compared	N	Cases Matched Correctly		Diff.	CR	P
		No.	Prop.			
Hospital patients	11	4	.36	$.05 \pm .27$	<1	
Outpatients	17	7	.41			
Psychotic patients	10	2	.20	$.30 \pm .14$	2.15	.03
Other NP patients	18	9	.50			
Total interviews:						
Over 19	14	5	.36	$.07 \pm .18$	<1	
Under 19	14	6	.43			
Interviews after testing:						
Over 11	14	6	.43	$.07 \pm .18$	<1	
Under 11	14	5	.36			
Frequency of interviews:						
Once weekly	15	6	.40	$.02 \pm .18$	<1	
Over once weekly	13	5	.38			
Cases 1-15	15	2	.13	$.56 \pm .16$	3.5	<.01
Cases 16-28	13	9	.69			
Check list first	14	7	.50	$.14 \pm .19$	<1	
Matching first	14	5	.36			

Only one of these differences, the order in which the Rorschach tests were administered—and interpreted—is significant at less than the 1 per cent level of confidence. A possible explanation of this difference is that the reports in the last half of the sample might have been more incisive descriptions than the first fifteen interpretations. The ratings of heterogeneity, which were made in the procedure for selecting the matching groups, provided some measure of the qualitative differences among the reports—at least within each of the two halves of the sample, but, unfortunately, not



over the entire sample. The results of this rating procedure, as shown in Table 2, indicate that in both halves of the sample, a significantly greater number of the comparisons were rated as different from one another (*O* or *OO*) than might be expected if the distribution of ratings had been even; on the other hand, the percentage of *S* and *SS* ratings was much less than expected. Thus, the efforts of the interpreter to achieve distinctive reports were sustained within each half of the samples. Whether or not this distinctiveness increased progressively from one half of the sample through the next cannot be stated conclusively inasmuch as not each interpretation was paired with every other interpretation throughout the whole sample. However, in view of the fact that no significant difference existed between the two halves of the sample in the proportion of *O* + *OO* ratings, it may be considered that the cases of the second half were no more distinctive, as compared among themselves, than those of the first half.

TABLE 2

Differences Between Obtained and Expected Ratings of *S* + *SS* and *O* + *OO*  
(Assuming an Expected Chance Distribution of Equal Proportions of *SS* + *S*,  
*SO* + *X* and *OO* + *O*.)

Cases	Comparisons	Rating	Obtained	Per Cent Obtained	Per Cent Expected	Diff.	CR	P
1-15	105	<i>S</i> + <i>SS</i>	16	15.2	33 + 4.47	-17.8	3.73	< .01
		<i>O</i> + <i>OO</i>	44	48.8		+10.8	2.06	< .05
16-28	78	<i>S</i> + <i>SS</i>	9	11.5	33 + 4.9	-21.5	4.39	< .01
		<i>O</i> + <i>OO</i>	38	48.7		+15.7	3.21	< .01

The second of these variables which showed a significant difference in the matching results was the diagnostic classification of the patients. Fewer of the cases diagnosed "psychotic" were matched correctly than those classed as "neurotic" or in other neuropsychiatric categories. Although the number of patients in these classifications was too small for further computation of differences, it was noted that seven of the ten psychotic cases fell in the first half of the sample. These results, if meaningful, would indicate that the interpretations of the Rorschach of the psychotic patients may have been less differentiating ones. In view of the fact that psychotic patients (other than paranoid types) often give vague and diffuse responses, it is to be expected that these interpretations would be less meaningful and distinctive. Such responses from the psychotic patients are consonant with the theory of personality used here, i.e., that inadequate perceptual differentiation is equated with psychoses. However, this concept is not helpful in distinguishing one psychotic patient from another, as was required of the matching

judges. If the judges operated on the basis of this concept also, then the criteria may have been as nondifferentiating as the Rorschach. Further study of the psychotic individual may be required, by both the Rorschach and other methods of observation, before a higher validity of interpretation can be demonstrated by the matching method.

#### *IV. The Check List Approach*

The check list approach consisted of the following steps:

1. A list of thirty-four multiple choice items was constructed, consisting of statements commonly used in interpretative reports and in psychotherapeutic analysis.

2. The reliability of the therapists in the use of this check list was determined, using a sample case analysis as criterion.

3. Four Rorschach interpreters independently checked their choices on the multiple choice items for the twenty-eight Rorschach protocols. The reliability of these Rorschach judges was determined by computing the significance of the number of agreements between these judges, for each item.

4. The therapists checked choices on each item on the basis of their impression of their patients. The validity of the Rorschach judges' choices was then determined by computing the significance of the number of times that they agreed with the therapists on each item.

##### A. SELECTION OF THE ITEMS

In order to obtain a list of multiple choice items representative of the many abstractions used in Rorschach interpretation, each of the major categories of personality utilized in the interpretative reports was represented by at least one item. These six major areas of personality were as follows: (a) inner drives and attitudes, (b) emotional reactions and relationships, (c) sensitivity to emotional stimuli, (d) intellectual functioning and reality testing, (e) sexual attitudes and identification, and (f) anxiety and defenses against anxiety. Each of these major categories or areas was further considered in four subdivisions or "dimensions": (a) the *frequency* or extent to which these areas were represented in his reaction to the test materials or to psychotherapy, (b) the characteristic *type* or nature of each area, (c) the *role* which each area played in the total pattern of the personality, (d) the *control* or manner in which the individual handled the attitude or reaction in question.

The questions asked in the interpretative reports concerning the indi-



vidual's attitudes toward his identity and his inner motivations were represented on the check list by those items referring to fantasy life and inner drives, as follows:

(Quantity) No. 17. "Expression by the individual of his inner needs and drives, i.e., his striving for satisfaction of these drives, is: almost completely absent," to "directly impulsive, showing an infantile lack of control."

(Role) No. 23. "Such inner fantasy life as the individual may allow himself is utilized for, or functions in his personality structure as: A. An internalization of certain unacceptable feelings, not permitted in overt behavior, e.g., for introjection of hostility in an intrapunitive manner. B. An attempt to organize and handle outer behavior in an integrated manner. C. A retreat from nearly all environmental frustrations, especially those in interpersonal relationships, with a handling of such relationships on a fantasy level. D. Very little, being poorly developed. E. Very little, being a source of anxiety in itself."

(Control) No. 13. "The method by which the individual handles and controls his inner emotional drives is chiefly: A. By fantasy solutions—possibly by divorcing such feelings from reality. B. By creative use of his energies, in a sublimated manner. C. By repressing them in a rigid and constricted manner. D. By direct release in overt behavior. E. By attempting to intellectualize, depersonalize or otherwise detach them from him emotionally."

Closely related to this general area of inner motivation are the individual's attitudes toward his sexual functioning, which were sampled on the check list by the following items:

(Quantity) No. 6. "The extent to which the individual enters into heterosexual relationships: is almost completely nil," to "is so exaggerated as to pervade much of the individual's behavior."

(Type) No. 30. "The following attitude may be considered as the 'basic' one with which the individual regards his own sexual or 'sexualized' behavior: A. As an aggressive (sadist) act. B. As a dangerous (castrating) act. C. As a passive, receptive (incorporative) act. D. As a demonstration of potency, an egotistic self-assertion, (autoerotic or exhibitionistic). E. As normal and acceptable (genital supremacy)." (A–D assume infantile sexual fixations or conflicts.)

No. 32. "The individual's general identification in most sexual and social roles is: A. With a dominant male figure. B. With a dominant female figure. C. With a passive male figure. D. With a passive female figure. E. Without a definite character and/or extremely ambivalent."

(Role) No. 10. "Homosexual relationships are utilized by the individual for, or play a role in his personality as: A. An integrated and mature part of his social behavior. B. A denial of rejection by other males. C. A denial of rejection by females. D. A minor role (e.g., for further satisfaction of narcissistic needs). E. An assertion of identification as to sexual role."

(Control) No. 24. "The chief method by which the individual handles his homosexual relationships is: A. By fairly overt emotional attachments possibly including sexual satisfactions. B. By repression of such feelings and/or avoidance of such relationships. C. By sublimating such feelings into socially acceptable channels of

behavior. D. By retreating into fantasy solutions. E. By intellectually detaching the emotional aspects, depersonalization."

No. 34. "The individual handles, or reacts to, possible heterosexual relationships (or needs for such relationships) chiefly by: A. A retreat into fantasy (without necessarily breaking with reality). B. By accepting social restrictions, and sublimating where necessary. C. By repressing such drives, and/or depersonalizing them, detaching the emotional aspects of such relationships. D. By affective outburst—such as overt anxiety and panic, etc. E. By breaking with reality."

The manner in which the individual perceives and accepts the pressure of his environment was represented in the following items, which dealt with interpersonal relationships and emotional reactions:

(Quantity) No. 27. "The degree to which the individual allows himself to become involved in emotional relationships with others is: a very limited involvement of any type," to "a purely volatile and explosive reaction."

(Type) No. 25. "The emotional tone or affect which the individual displays in his emotional attachments with others is most often: A. Warm and spontaneous. B. (Absent.) C. Cold and detached. D. Hostile and oppositional. E. Forced and artificial.

(Role) 33. "Involvement in active emotional, interpersonal relationships serves the individual as, or plays a role in his personality structure as: A. A method of compensating for the inadequacies felt within himself. B. A release mechanism for the satisfaction of inner drives. C. A minor role—in an intraversive adjustment, under accumulated frustration or increasing environmental stimulation only. D. (None.) E. As a mature and integrated part of his behavior."

(Control) No. 19. "The principal method by which the individual handles and controls his emotional reactions in his interpersonal relationships is: A. By integrating them in a mature manner with other personal needs. B. By rigidly avoiding and denying the emotional aspects (isolation of emotion). C. By ignoring the reality of such an emotion and autistically withdrawing into fantasy solutions. D. By immature, and possibly aggressive, reactions. E. By depersonalizing such situations through intellectualizing or rationalizing.

Four other items were constructed requiring judgment relative to the individual's sensitivity to environmental stimulation:

(Quantity) No. 8. "The extent to which the individual allows himself to be receptive to the affective feelings of others, or to other emotional stimulation: A. Is limited and tenuous, chiefly when socially approved. B. Is practically absent. C. Is such that the individual is acutely aware of the emotional aspects of a situation. D. Shows a well-balanced and integrated sensitivity and tact. E. Shows a tendency to be unduly sensitive."

(Type) No. 9. "The individual's most characteristic reaction to the affective feelings of others and to the emotional stimulation from his environment, is: to be indifferent and disinterested," to "to be overtly sensuous."

(Role) No. 16. "Sensitivity to the emotions of others or to other emotional stimuli is utilized by the individual for, or plays a role in his personality as: A. (None.) B. A counterreaction to repressed hostility. C. A withdrawal from frustra-



tion, from a more active emotional involvement. D. An integrated part of a mature handling of social relationships. E. A primary source of guilt and anxiety."

(Control) No. 2. "The way in which the individual controls and handles possible sensitivity to emotional stimulation, especially to the feelings of others, is usually: A. By rigid repression and restriction of such sensitivity. B. By attributing such stimulation to his own inner needs (by introjection). C. By reactively denying such feelings in aggressive, hostile behavior. D. By divorcing such feelings from reality and/or by fantasy solutions. F. By acceptance and integration of such sensitivity in social relationships.

From the area of intellectual functioning and of reality testing, the following items were derived:

(Quantity) No. 1. "The wealth of the individual's intellectual activity may be characterized as generally: impoverished, and tending to be perserverative," to "having a wide range of interests, often being rich and original in content." No. 7. "The individual's intellectual productivity may be estimated as generally: very limited," to "extensive."

(Type) No. 15. "The individual's intellectual approach to a problem or situation: A. Usually shows a tendency to abstract and overgeneralize, without sufficient attention to everyday affairs. B. Tends to be overly critical, analytical, possibly picayunish. C. Usually shows a fair ability to conceptualize, but with adequate attention to practical concrete matters. D. May contain some evidence of delusional thought processes, forcing relationships between facts or distorting reality. E. Is most often a matter-of-fact approach, tending to be overly concrete."

No. 18. "The individual's ties to reality may be classified as chiefly: very strong, as never permitting any vagueness" through "adequate—but not overly concerned with reality testing," through "so tenuous as to easily become inadequate" to "quite inadequate and/or absent."

(Role) No. 22. "Intellectual functioning is utilized by the individual for, or has a principal function in his personality structure as: A. A rigid defense against the release of inner drives and/or emotional ties with others, by depriving them of their emotional tone. B. A mature and normal mode of controlling himself and his environment. C. As an aid to autistic thinking, e.g., in delusional types of solutions. D. A highly aggressive, critical defense mechanism. E. Only a minor role, e.g., as an aid to immediate satisfaction of narcissistic needs."

(Control) No. 5. "The individual's contact with reality appears weakest: A. In his creative, inner fantasy life. B. In his active, and potentially aggressive, interpersonal relationships. C. In his sensitivity to emotional stimulation. D. In seemingly impersonal situations (to which affect has been displaced). E. In his release of instinctual drives."

The items included on the check list under the rubric of *anxiety* correspond in part to those considerations given ego-functioning above:

(Quantity) No. 28. "The degree to which the individual shows feelings of generalized disturbance may be estimated as: seldom more than a minimal and occasional uneasiness," to "states of overwhelming panic."

(Type) No. 3. "The individual's expression of feelings of generalized disturbance may be characterized as: A. A free-floating type of anxiety state. B. A feeling of inner tension and conflict-guilt feelings. C. Overt depression. D. A sense of frustration and disappointment. E. (Relatively absent.)"

(Role) No. 21. "The effect of anxiety and/or guilt feelings on the individual's personality structure constitutes: no noticeable effect," to "a gross breakdown of most functioning."

The different defenses were considered more specifically in the following items:

(Projection) No. 4. "Projection of guilt feelings onto others or the environment is used by the individual as a method of averting anxiety to the following degrees: very rarely," to "extensively."

(Rationalization) No. 11. "The individual uses rationalization and justification as an intellectual evasion of anxiety to the following degree: very rarely," to "extensively."

(Obsession) No. 12. "The individual uses compulsive behavior or obsessive thinking as a magical ritualistic denial of anxiety to the following degree: very rarely," to "extensively."

(Displacement) No. 14. "The individual attempts to avert anxiety by displacement of emotional content to some more 'neutral' situation: very rarely," to "extensively."

(Withdrawal) No. 20. "The individual attempts to avert anxiety by fantasy solutions and/or by withdrawal from contact with reality to the following degree: very rarely," to "extensively."

(Normal reaction) No. 26. "The individual uses anxiety as a normal 'warning signal' of possible frustration: very rarely," to "extensively."

(Acting out) No. 29. "The individual attempts to avert anxiety by 'acting out' of frustrations onto the environment, by negativism and aggression, etc.: very rarely," to "extensively."

(Isolation) No. 31. "The individual attempts to avert anxiety by rigid isolation of all emotional aspects of a situation: very rarely," to "extensively."

#### B. ARRANGEMENT OF THE ITEMS ON THE LIST

It is possible that, if the items had been presented in the logical order of the scheme shown above, a judge's choice on one item might directly influence his choices on succeeding items, especially those within the same major category. In order to lessen the sequential effect, the items were presented in a random arrangement.

#### C. THE NUMBER AND ORDER OF THE CHOICES

For the sake of uniformity, five choices were listed for each statement. As was discovered afterward, this uniform number of choices was an unnecessary restriction; in many instances, a larger number of choices would have



offered the judges more opportunity to describe their patients, and in a more accurate manner. In many instances, also, these five choices formed an obvious continuum, e.g., from "extensively" to "rarely," or "well adjusted" to "very disturbed," etc. Since a judge's choices on one item might well be influenced by the position on this continuum of his choices on previous items, the order of choices was varied in a random manner from item to item.

#### D. THE INSTRUCTIONS

The judges were instructed to make two choices on each item for each patient. Second choices were requested, because (a) it was possible that a pair of judges might agree, given two choices each, even though they might not agree on a first choice; and (b) some judges felt less forced in their judgments when allowed two choices.

#### E. RELIABILITY OF THE THERAPISTS IN THE USE OF THE CHECK LIST

Reliability of the therapists in the use of the check list was determined in the same manner as was their reliability in the matching procedure. Using a sample case analysis (described above) as a criterion, ten of the therapists indicated choices on each item of the list. The therapists had previously had some training in the use of this type of list, in a trial run of a preliminary form. A rough estimate of the reliability of these therapists was obtained by assuming the agreement to be satisfactory when five or more therapists indicated the same first choice on an item—as describing this sample case history. When any two judges indicated the same two choices on any item, either as first or second choice, this was also noted as an agreement. For these "pairs" of choices, satisfactory agreement on an item was assumed when four or more judges used the same pair. On twenty of the items, five or more judges employed the same pair of choices. Although this reliability study was admittedly limited, it seems reasonable to assume that similar results might have been obtained if a more extensive study had been possible. Strictly speaking, the results of this step in the check list approach can only be taken to indicate that the therapists were able to agree satisfactorily on a majority of the items, using one sample case as a basis of judgment.

#### F. RELIABILITY OF THE RORSCHACH JUDGES IN THE USE OF THE CHECK LIST

The twenty-eight Rorschach records of the sample were judged independently on the check list by four experienced Rorschach interpreters.

These judges<sup>9</sup> received a brief training in the use of this check list on a sample Rorschach protocol. The reliability of these Rorschach judges in the use of the check list was considered in terms of the number of agreements, i.e., the number of times they indicated the same choice on an item regarding the same individual. For each item, the twenty-eight first choices of each judge were compared with those of every other judge—two judges at a time. Thus, six sets of comparisons were made for each item. The first choices of each pair of judges were tabulated on a five-square table, such that the agreements fell on the diagonal. The degree of agreement expected by chance was then computed on the assumption that the two sets of judgments were independent, subject to the restriction of the observed marginal totals. Since in most instances this chance degree of agreement was a relatively small proportion of the total number of cases—usually less than one-third—the significance of the difference between this number of agreements expected by chance and the number of agreements actually obtained was again read from the tables of Poisson's distribution (14). Where the expected number of agreements was larger than one-third of the total number of cases, the significance of this difference was computed in terms of the standard error of a proportion.

All in all, the four judges of the Rorschach agreed significantly about one-third of the time. Of the 204 comparisons for first choices (the choices of six pairs of judges compared on 34 items), 78 resulted in agreement at the 10 per cent level or beyond; 51 of these were significant at the 5 per cent or beyond; and 26 at the 2 per cent level or better. For the two choices combined, 63 of the obtained agreements were significant at the 10 per cent level or beyond; 43 of these at the 5 per cent level; and 31 at the 2 per cent level or better. If we consider the rather abstruse wording of some of these items and the limitations on the number of choices, this degree of agreement is considerable. The large percentage of lack of agreement is not surprising in view of the fact that the judges were attempting to make almost *unqualified* statements about personality from such a restricted sampling of behavior, i.e., responses to ten ink-blot pictures. The alternative or second choice did not appreciably increase the agreement between the judges.

Further analysis of the agreement and lack of agreement indicates that a fair number of items may be called reliable. If three significant agreements, i.e., agreement by three pairs of judges, be granted as indicating satisfactory reliability for an item, then about one-third of the 34 items may be called reliable—12 for first choices and 11 for the combined choices. An appreci-

9. Mr. Walter Klopfer, Mr. J. Neil Campion, Jr., and Dr. Claire Thompson of the University of California were kind enough to devote many hours assisting the author in making these judgments.



able part of the lack of agreement may have been attributable to some particular pair of judges. However, it was found that all pairs of judges agreed to about the same degree. The degree of agreement within each area and within each dimension, as shown in Table 3, is summarized as a *proportion of the total number of comparisons* made in that area or dimension which showed significant agreement. The area in which the proportion of agreement was highest was intellectual functioning (.667). This particular area has been given close attention in the interpretative method of Klopfer and Kelley (10). Besides, it has also been thoroughly discussed in the current literature; therefore, the comparatively strong agreement shown here is not surprising. Two items which were included in this area, but which failed to have three or more agreements (Nos. 5 and 18) should properly belong to reality testing in general, rather than to specific intellectual functioning. Undoubtedly, a better definition of *reality testing* should have been reached by the judges.

The judges agreed reasonably well on another area: the one dealing with the individual's inner life, including his motivations and incorporated attitudes (.611). The significance of this agreement lies in the fact that this area is not as easily interpreted as that of intellectual functioning. This result might be explained by the hypothesis that the Rorschach is designed to tap these inner drives more than it does other areas. However, the degree of agreement about these inner drives does not appear to be much stronger than that about outer emotional reactions (.417) or sensitivity (.458). (The actual number of agreements was too small to permit statistical comparison.)

The lower percentage of agreement regarding emotional reactions and sensitivity may be traced to the disagreement among the judges about the dimensions of *type* and *role*. These two dimensions in these areas were the most difficult to restrict to five choices, and it seemed that greater agreement might have occurred if more choices had been provided.

Although the agreement on items concerning sexual attitudes (.277) might also have been improved by expanding the number of choices, the basic difficulty was more likely the confusion over the concepts employed in these items, particularly the concept "homosexual." This term was intended to refer to any type of attitude toward the same sex, rather than just to overt sexual behavior. Although the judges understood this broader connotation, they tended to limit their thinking about it to the latter, more common meaning.

The notably low proportion of agreement on the items pertaining to anxiety and defenses (.166) (see Table 4) is again explicable by the fact of the limited number of choices. To paraphrase: all anxiety is scarcely divisible into five parts! A more serious source of disagreement was found to be

TABLE 3  
Distribution of Significant Agreement Between Rorschach Interpreters According to  
Area and Dimension of Personality Functioning

Area	No. of Comparisons	Proportion of Agreement	Frequency		Type		Role		Control	
			Items	No. of Agree- ments	Items	No. of Agree- ments	Items	No. of Agree- ments	Items	No. of Agree- ments
Inner drives	18	.611	17	2	—	—	23	4	13	5
Sexual attitudes	36	.277	6	0	30	3	10	0	34	4
					32	1			24	2
Emotional reactions	24	.417	27	4	25	2	33	1	19	3
Sensitivity	24	.458	8	3	9	1	16	1	2	6
Reality testing	36	.667	1	6	15	6	22	3	5	2
			7	5	18	2				
Anxiety reactions	18	.166	28	2	3	1	21	0	—	—
Number of comparisons			42		42		36		36	
Proportion of agreement			.524		.328		.250		.606	



the poor definition of terms in this area. Even after the use of the check list had been reviewed among the judges, such terms as "displacement" were not consistently applied; note the significant disagreements on this item (No. 14) particularly. Several of the judges expressed the opinion that the items referring to defenses often overlapped in meaning or were otherwise confusing.

TABLE 4  
Distribution of Significant Agreements Between Rorschach Interpreters,  
According to Type of Defense

Defense	Item No.	No. of Agreements
Projection	4	2
Rationalization	11	0
Obsession	12	1
Displacement	14	-2
Withdrawal	20	0
Normal warning	26	2
Acting out	29	1
Isolation	31	2
Total no. of comparisons		48
Proportion of agreement		.166

Viewing the results according to *dimension*, the highest agreement occurred on those items concerning the individual's handling or *control* of his functioning (.606). Next on the scale of agreement came the dimension of *frequency* (.524). These results indicate that Rorschach interpreters can agree among themselves on the two following aspects: the way in which an individual controls his impulses, and the degree to which he uses any particular control or defense or expresses an attitude.

The difference in the percentage of agreement between *control* (.606) and *role* (.250) assumes an importance when one considers that both dimensions deal with *relationships* between reactions. Thus, the judges were in much better agreement as to the *control* relationship between various areas of personality than they were concerning the relative importance and relationship of a particular reaction or attitude in the individual's overall functioning. This lack of agreement about the concept of *role* may be due partially to inadequate phrasing of the items or limitations in the number of choices. But the fact should also be pointed out that Rorschach interpretation underscores this concept of *control*, and that, too often, little attention is given the importance of such control or defense in the "economics" of the personality.

Thus, if it be granted that Rorschach theory emphasizes the perceptual functioning of the ego, then the comparatively higher reliability on the items covering *intellectual functioning* and *control* was to be expected. Conversely, these results may be taken to show a lack of agreement on those aspects which are less clearly defined in Rorschach theory, namely, the contentual factors, such as sexual attitudes and types of anxiety.

#### G. RESULTS OF VALIDATION ON THE CHECK LIST

The choices on each item by each of the four Rorschach judges were tabulated with those of the therapists—one pair of judges at a time; four differences between the obtained frequencies of agreement and that expected by chance were derived for first choices, and the significance of these differences was noted in terms of the approximation to binomial probabilities read from Poisson's tables.

Considering first choices alone, only nine of the 136 comparisons (four comparisons on the thirty-four items) resulted in significant differences, at the 10 per cent level of confidence or beyond. No significant agreement appeared for the combined choices (first and second choices together). Only five of the differences on first choices were "positive differences," i.e., the obtained agreement was larger than that expected by chance, while four were "negative differences," i.e., the obtained agreement being smaller than chance. Since, in a distribution of 136 differences, 13.6 might be expected to show such significance by chance, this small number of significant agreements cannot be considered to be of any statistical importance. The differences did not seem to occur in any meaningful pattern: no more than one difference occurred on any one item; almost an equal number of such differences occurred for each pair of judges; and these differences did not appear to have any relation to the scheme of personality areas and dimensions.

In general, these results support the hypothesis that the check list type of approach is not applicable to the validation of the interpretations of projective techniques when these techniques are interpreted by means of a dynamic concept of personality. Such statements as those used in this check list are apparently meaningless, except in the context of an integrated descriptive report.

Undoubtedly, the amount of agreement between the Rorschach judges and the therapists was dependent on the degree to which these two sets of judges agreed separately among themselves. The reliability studies discussed above showed that when one set of judges used one kind of data, either Rorschach records or therapy, the agreement in that instance was satis-



factory. In terms of the number of items which showed significant agreement, the Rorschach judges agreed among themselves 35 per cent of the time, i.e., on twelve items, while the therapists agreed among themselves fifty-eight per cent of the time, or on twenty items. In view of these reliabilities, there would seem to be some chance of obtaining higher validity. At the same time, one might inquire as to why these isolated statements were not applicable to the *validation* of Rorschach interpretation when they appear to be applicable to the study of its *reliability*.

Qualitative examination of the reliability of the Rorschach judges and of the therapists may serve to explain why there was agreement within each respective set of judges but little agreement between the two sets. In fact, the results of this reliability study may aid in the exploration of the relationships between the isolated statements and the whole reports.

As has been noted in the above discussion, the Rorschach judges agreed among themselves on those statements referring to the concepts which are most clearly defined in Rorschach theory, namely, in the area of *intellectual functioning* and on the dimension of *control*. Although several other basic concepts are commonly recognized in Rorschach interpretation, these two concepts may be considered as the "axis" from which the whole pattern of the personality is evolved. It must be acknowledged that when other concepts are introduced in an interpretation, they are not based entirely on the relationship between those two primary inferences. However, the analysis of clues in the Rorschach protocol associated with other concepts than intellectual functioning or control is strongly influenced by the conclusions about these two concepts.

Although the therapists used the same general framework of concepts as adopted by the Rorschach interpreters, there was no direct evidence as to which particular concepts within this framework were central in each therapist's thinking. Since the therapy dealt principally with emotional relationships and with the analysis of emotional reactions, one might expect that the concepts concerned with this area of the personality determined the therapist's orientation. Thus, in considering a patient's personality as observed during therapy, the therapist might be inclined to give weight only to such intellectual functioning as was directly related to the patient's emotional life.

Although the Rorschach interpreter and the therapist may have had different concepts and clues in mind as they analyzed their respective observations of the patient's reactions, they were both attempting to infer a total picture of the individual's functioning, i.e., his personality structure. Each set of judges used their particular clues and concepts consistently and reliably when considering their respective data. The two sets of judges did

not agree significantly on the concepts which were not central to their separate considerations, especially when these concepts were isolated from the context of the whole structural pattern—as was required on the check list. On the other hand, when the total picture of the individual was taken into consideration, agreement between the therapists and the Rorschach interpreter was obtained, as was demonstrated in the matching experiment.

### *V. The Relationship Between the Two Approaches*

We return to the hypothesis which assumed that validity of (or agreements on) specific statements is more likely to be found when the whole descriptive report is validated. Bearing in mind that of the twenty-eight whole reports in this study, only eleven were satisfactorily matched or validated, then little or no validity could be expected throughout the check list *for the entire sample*. On the other hand, more agreement would be expected on the check list items for those eleven cases on whom there was also agreement on the whole reports. In order to test this hypothesis, the sample was divided into two groups, based on the results of the matching experiments, i.e., the eleven cases correctly matched, and the seventeen cases in which matching was unsuccessful.

As one test of the hypothesis, the significance of the agreements between the Rorschach and criterion judges was calculated separately for these two groups on each item of the check list. This division yielded no significant agreement for any one of the items for either group, whether considering the first choices separately or the combined choices. These results were not surprising, considering the results of the validation of these items for the sample as a whole. In addition, a positive agreement significantly above chance was contraindicated because of the small number of cases in each of the two groups.

A preliminary analysis of the data indicated that the obtained agreements on the check list for the eleven matched cases were consistently greater than those for the other seventeen cases, although not enough to be statistically significant. This possible difference was again tested by contrasting the obtained agreements and disagreements for both groups on fourfold chi-square tables. Since the number of agreements on first choices was too small to permit this comparison for any item, a chi-square test was made on the first and second choices combined. No significant results were obtained on the whole. Only eleven of the 136 comparisons attained significance beyond the 10 per cent level of confidence. Five of these chi-squares indicated



a difference in the expected directions, i.e., the correctly matched group had more agreements but six were in the opposite direction, i.e., the group not correctly matched possessed greater agreement on the check list item in question than did the cases correctly matched. There was no indication that these results were associated with any particular pair of judges or with any definite area of content of the items.

The absence of *any* significant degree of difference between these two groups of cases indicates that the agreement between the Rorschach and therapy judges on the specific items had no relation to their agreement on the whole reports. The lack of significant agreement between the Rorschach judges and the therapists on the check list, for the whole sample, therefore, cannot be attributed to the presence of any group of cases in the sample which were not validly interpreted as whole reports.

## *VI. Summary*

### **A. THE CHECK LIST APPROACH**

The check list approach was employed in this study for two purposes: (a) to study the possibilities of validating a projective technique on a set of isolated interpretative statements, and (b) to determine whether the behavior of the individual on the test and in a life situation could be described by the same statement. The results of validating an accepted projective technique, the Rorschach, on this check list indicate that this approach is not applicable to the validation of such projective techniques. In view of the fact that both the Rorschach interpreters and the therapists used this check list reliably when describing their respective observations, the main conclusion is that the check list approach may be applicable to the study of personality descriptions only when a common set of concepts is maintained as a reference point for inferring the total pattern of the individual's functioning.

In general, these findings support the contentions of Vernon (19), Frank (3), and other investigators who have argued that a description of the interrelated functioning of an individual can be validated only as a whole. In particular, it has been demonstrated that a total and integrated picture of the individual's personality may be valid, even though there may be no more than chance agreement between the judgments of a test interpreter and a criterion in regard to isolated statements about the individual's functioning.

This check list approach was a more rigid and exacting test of the validity of separate interpretative statements than the "item analysis" design

employed by Harrison (4) and by Cronbach (2). In the first place, neither of these investigators attempted to test whether the behavior of the individual in both the test situation and in the life situation could be described by a common set of statements. Secondly, in this check list approach, the isolated statements were selected according to a definite rationale, i.e., the scheme of "area" and "dimensions" of personality.

The results of the present study indicate that the behavior of the individual in both the test situation and a life situation could not be satisfactorily described by the same statement. Since the "item analysis" method does not make this demand on the test, it is probably a sounder approach. However, in the use of an "item analysis" approach it is recommended that the rationale for selecting statements from the whole reports be clearly stated.

#### B. THE MATCHING APPROACH

The chief advantage of the matching approach, according to Vernon (18), is that for whole interpretative reports, this approach tests the validity of the most essential features of the interpretation of projective techniques, i.e., the accuracy with which the *interrelated pattern* of the individual's functioning is described. The findings of the present study support Vernon's contention. By means of the matching approach, validity was demonstrated for descriptions of personality which emphasized these interrelationships. On the other hand, no validity was obtained when isolated interpretative statements were applied in which the relationships between various functionings of the personality were not elaborated.

One of the major concerns in the present study of the matching approach was the effect of the heterogeneity of the matching groups. This study indicated that use of matching groups of an "optimum" heterogeneity resulted in a smaller number of successful matchings than were obtained in previous studies using randomly selected groups. Since the matching approach is essentially a test of *interindividual* differentiation, careful attention must be paid to the nature of the sample of individuals from whom a particular individual is to be differentiated.

The chief criticism directed against the matching approach has been that it does not provide a test of the accuracy with which various part-functionings within this whole pattern are delineated. The present study attempted, without success, to test this *intra-individual* differentiation of an interpretation, by means of the check list approach. Since the completion of the present study, an "item analysis" design has been suggested by Cronbach (2) which, if used in conjunction with a matching approach, may provide a thorough statistical method for the validation of projective techniques.



## REFERENCES

1. CHAPMAN, D. W. The statistics of correct matching. *Amer. J. Psychol.*, 1934, 46, 287-298.
2. CRONBACH, L. J. A validation design for qualitative studies of personality. *J. Consult. Psychol.*, 1948, 12, 365-375.
3. FRANK, L. K. Projective methods for the study of personality. *J. Psychol.*, 1939, 8, 398-413.
4. HARRISON, R. Studies in the use and validity of the TAT with mentally disordered patients. II. A quantitative validity study by the method of blind analysis. *Character & Pers.*, 1940, 9, 122-138.
5. HERTZ, MARGUERITE R. The validity of the Rorschach method. *Amer. J. Orthopsychiat.*, 1941, 11, 512-520.
6. HERTZ, MARGUERITE R. Rorschach: twenty years after. *Psychol. Bull.*, 1942, 39, 529-572.
7. HERTZ, MARGUERITE R. The Rorschach method: Science or mystery? *J. Consult. Psychol.*, 1943, 7, 67-79.
8. HERTZ, MARGUERITE R., AND RUBENSTEIN, B. A comparison of three blind Rorschach analyses. *Amer. J. Orthopsychiat.*, 1939, 9, 295-315.
9. HUNTER, M. E. The practical value of the Rorschach test in a psychological clinic. *Amer. J. Orthopsychiat.*, 1939, 9, 278-294.
10. KLOPPER, B., AND KELLEY, D. M. *The Rorschach technique*. New York: World Book Co., 1942.
11. KRUGMAN, JUDITH I. A clinical validation of the Rorschach with problem children. *Rorschach Res. Exch.*, 1942, 5, 61-70.
12. MACFARLANE, JEAN W. Problems of validation inherent in projective methods. *Amer. J. Orthopsychiat.*, 1942, 12, 405-410.
13. MURRAY, H. A., AND ASSOCIATES. *Explorations in personality: A clinical and experimental study of fifty men of college age*. New York: Oxford Univ. Press, 1938.
14. PEARSON, K. *Tables for statisticians and biometricians*. Cambridge, Eng.: Cambridge Univ. Press, 1914.
15. ROSENZWEIG, S. An outline of a cooperative project for validating the Rorschach test. *Amer. J. Orthopsychiat.*, 1935, 5, 395-403.
16. SYMONDS, P. M., & KRUGMAN, M. Projective methods in the study of personality. *Rev. Educ. Res.*, 1944, 14, 81-93.
17. TOMKINS, S. S. *The Thematic Apperception Test*. New York: Grune and Stratton, 1947.
18. VERNON, P. E. The significance of the Rorschach test. *Brit. J. Med. Psychol.*, 1935, 15, 199-217.
19. VERNON, P. E. Matching methods as applied to the investigation of personality. *Psychol. Bull.*, 1936, 33, 149-177.

Lee J. Cronbach

Paul E. Meehl<sup>1</sup>

## CONSTRUCT VALIDITY IN PSYCHOLOGICAL TESTS

**V**ALIDATION OF psychological tests has not yet been adequately conceptualized, as the APA Committee on Psychological Tests learned when it undertook (1950-54) to specify what qualities should be investigated before a test is published. In order to make coherent recommendations the Committee found it necessary to distinguish four types of validity, established by different types of research and requiring different interpretation. The chief innovation in the Committee's report was the term *construct validity*.<sup>2</sup> This idea was first formulated by a subcommittee (Meehl and R. C. Challman) studying how proposed recommendations would apply to projective techniques, and later modified and clarified by the entire Committee (Bordin, Challman, Conrad, Humphreys, Super, and the present writers). The statements agreed upon by the Committee (and by committees

Reprinted from *Psychol. Bull.*, 1955, 52, 281-302, by permission of the American Psychological Association and the authors.

1. The second author worked on this problem in connection with his appointment to the Minnesota Center for Philosophy of Science. We are indebted to the other members of the Center (Herbert Feigl, Michael Scriven, Wilfrid Sellers), and to D. L. Thistlethwaite of the University of Illinois, for their major contributions to our thinking and their suggestions for improving this paper.

2. Referred to in a preliminary report (58) as *congruent validity*.



of two other associations) were published in the *Technical Recommendations* (59). The present interpretation of construct validity is not "official" and deals with some areas where the Committee would probably not be unanimous. The present writers are solely responsible for this attempt to explain the concept and elaborate its implications.

Identification of construct validity was not an isolated development. Writers on validity during the preceding decade had shown a great deal of dissatisfaction with conventional notions of validity, and introduced new terms and ideas, but the resulting aggregation of types of validity seems only to have stirred the muddy waters. Portions of the distinctions we shall discuss are implicit in Jenkins' paper, "Validity for what?" (33), Gulliksen's "Intrinsic validity" (27), Goodenough's distinction between tests as "signs" and "samples" (22), Cronbach's separation of "logical" and "empirical" validity (11), Guilford's "factorial validity" (25), and Mosier's papers on "face validity" and "validity generalization" (49, 50). Helen Peak (52) comes close to an explicit statement of construct validity as we shall present it.

### *Four Types of Validation*

The categories into which the *Recommendations* divide validity studies are: predictive validity, concurrent validity, content validity, and construct validity. The first two of these may be considered together as *criterion-oriented* validation procedures.

The pattern of a criterion-oriented study is familiar. The investigator is primarily interested in some criterion which he wishes to predict. He administers the test, obtains an independent criterion measure on the same subjects, and computes a correlation. If the criterion is obtained some time after the test is given, he is studying *predictive validity*. If the test score and criterion score are determined at essentially the same time, he is studying *concurrent validity*. Concurrent validity is studied when one test is proposed as a substitute for another (for example, when a multiple-choice form of spelling test is substituted for taking dictation), or a test is shown to correlate with some contemporary criterion (e.g., psychiatric diagnosis).

*Content validity* is established by showing that the test items are a sample of a universe in which the investigator is interested. Content validity is ordinarily to be established deductively, by defining a universe of items and sampling systematically within this universe to establish the test.

*Construct validation* is involved whenever a test is to be interpreted as a measure of some attribute or quality which is not "operationally defined." The problem faced by the investigator is, "What constructs account for

variance in test performance?" Construct validity calls for no new scientific approach. Much current research on tests of personality (9) is construct validation, usually without the benefit of a clear formulation of this process.

Construct validity is not to be identified solely by particular investigative procedures, but by the orientation of the investigator. Criterion-oriented validity, as Bechtoldt emphasizes (3, p. 1245), "involves the *acceptance* of a set of operations as an adequate definition of whatever is to be measured." When an investigator believes that no criterion available to him is fully valid, he perforce becomes interested in construct validity because this is the only way to avoid the "infinite frustration" of relating every criterion to some more ultimate standard (21). In content validation, *acceptance* of the universe of content as defining the variable to be measured is essential. Construct validity must be investigated whenever no criterion or universe of content is accepted as entirely adequate to define the quality to be measured. Determining what psychological constructs account for test performance is desirable for almost any test. Thus, although the MMPI was originally established on the basis of empirical discrimination between patient groups and so-called normals (concurrent validity), continuing research has tried to provide a basis for describing the personality associated with each score pattern. Such interpretations permit the clinician to predict performance with respect to criteria which have not yet been employed in empirical validation studies (cf. 46, pp. 49-50, 110-111).

We can distinguish among the four types of validity by noting that each involves a different emphasis on the criterion. In predictive or concurrent validity, the criterion behavior is of concern to the tester, and he may have no concern whatsoever with the type of behavior exhibited in the test. (An employer does not care if a worker can manipulate blocks, but the score on the block test may predict something he cares about.) Content validity is studied when the tester is concerned with the type of behavior involved in the test performance. Indeed, if the test is a work sample, the behavior represented in the test may be an end in itself. Construct validity is ordinarily studied when the tester has no definite criterion measure of the quality with which he is concerned, and must use indirect measures. Here the trait or quality underlying the test is of central importance, rather than either the test behavior or the scores on the criteria (59, p. 14).

Construct validation is important at times for every sort of psychological test: aptitude, achievement, interests, and so on. Thurstone's statement is interesting in this connection:

In the field of intelligence tests, it used to be common to define validity as the correlation between a test score and some outside criterion. We have reached a stage of sophistication where the test-criterion correlation is too coarse. It is obsolete. If we attempted to ascertain the validity of a test for the second space-factor, for example, we would have to get judges [to] make reliable judgments about people as to this factor. Ordinarily their [the available judges'] ratings would be of no value as a



criterion. Consequently, validity studies in the cognitive functions now depend on criteria of internal consistency . . . (60, p. 3).

Construct validity would be involved in answering such questions as: To what extent is this test of intelligence culture-free? Does this test of "interpretation of data" measure reading ability, quantitative reasoning, or response sets? How does a person with A in Strong Accountant, and B in Strong CPA, differ from a person who has these scores reversed?

*Example of construct validation procedure.* Suppose measure X correlates .50 with Y, the amount of palmar sweating induced when we tell a student that he has failed a Psychology I exam. Predictive validity of X for Y is adequately described by the coefficient, and a statement of the experimental and sampling conditions. If someone were to ask, "Isn't there perhaps another way to interpret this correlation?" or "What other kinds of evidence can you bring to support your interpretation?" we would hardly understand what he was asking because no interpretation has been made. These questions become relevant when the correlation is advanced as evidence that "test X measures anxiety proneness." Alternative interpretations are possible; e.g., perhaps the test measures "academic aspiration," in which case we will expect different results if we induce palmar sweating by economic threat. It is then reasonable to inquire about other *kinds* of evidence.

Add these facts from further studies: Test X correlates .45 with fraternity brothers' ratings on "tenseness." Test X correlates .55 with amount of intellectual inefficiency induced by painful electric shock, and .68 with the Taylor Anxiety scale. Mean X score decreases among four diagnosed groups in this order: anxiety state, reactive depression, "normal," and psychopathic personality. And palmar sweat under threat of failure in Psychology I correlates .60 with threat of failure in mathematics. Negative results eliminate competing explanations of the X score; thus, findings of negligible correlations between X and social class, vocational aim, and value-orientation make it fairly safe to reject the suggestion that X measures "academic aspiration." We can have substantial confidence that X does measure anxiety proneness if the current theory of anxiety can embrace the variates which yield positive correlations, and does not predict correlations where we found none.

### *Kinds of Constructs*

At this point we should indicate summarily what we mean by a construct, recognizing that much of the remainder of the paper deals with this question. A construct is some postulated attribute of people, assumed to be reflected in test performance. In test validation the attribute about which

we make statements in interpreting a test is a construct. We expect a person at any time to possess or not possess a qualitative attribute (amnesia) or structure, or to possess some degree of a quantitative attribute (cheerfulness). A construct has certain associated meanings carried in statements of this general character: Persons who possess this attribute will, in situation X, act in manner Y (with a stated probability). The logic of construct validation is invoked whether the construct is highly systematized or loose, used in ramified theory or a few simple propositions, used in absolute propositions or probability statements. We seek to specify how one is to defend a proposed interpretation of a test; *we are not recommending any one type of interpretation.*

The constructs in which tests are to be interpreted are certainly not likely to be physiological. Most often they will be traits such as "latent hostility" or "variable in mood," or descriptions in terms of an educational objective, as "ability to plan experiments." For the benefit of readers who may have been influenced by certain eisegeses of MacCorquodale and Meehl (40), let us here emphasize: Whether or not an interpretation of a test's properties or relations involves questions of construct validity is to be decided by examining the entire body of evidence offered, together with what is asserted about the test in the context of this evidence. Proposed identifications of constructs allegedly measured by the test with constructs of other sciences (e.g., genetics, neuroanatomy; biochemistry) make up only *one* class of construct-validity claims, and a rather minor one at present. Space does not permit full analysis of the relation of the present paper to the MacCorquodale-Meehl distinction between hypothetical constructs and intervening variables. The philosophy of science pertinent to the present paper is set forth later in the section entitled, "The nomological network."

### *The Relation of Constructs to "Criteria"*

#### CRITICAL VIEW OF THE CRITERION IMPLIED

An unquestionable criterion may be found in a practical operation, or may be established as a consequence of an operational definition. Typically, however, the psychologist is unwilling to use the directly operational approach because he is interested in building theory about a generalized construct. A theorist trying to relate behavior to "hunger" almost certainly invests that term with meanings other than the operation "elapsed-time-since-feeding." If he is concerned with hunger as a tissue need, he will not accept time lapse as *equivalent* to his construct because it fails to consider, among other things, energy expenditure of the animal.



In some situations the criterion is no more valid than the test. Suppose, for example, that we want to know if counting the dots on Bender-Gestalt figure five indicates "compulsive rigidity," and take psychiatric ratings on this trait as a criterion. Even a conventional report on the resulting correlation will say something about the extent and intensity of the psychiatrist's contacts and should describe his qualifications (e.g., diplomat status? analyzed?).

Why report these facts? Because data are needed to indicate whether the criterion is any good. "Compulsive rigidity" is not really intended to mean "social stimulus value to psychiatrists." The implied trait involves a range of behavior-dispositions which may be very imperfectly sampled by the psychiatrist. Suppose dot-counting does not occur in a particular patient and yet we find that the psychiatrist has rated him as "rigid." When questioned the psychiatrist tells us that the patient was a rather easy, free-wheeling sort; however, the patient *did* lean over to straighten out a skewed desk blotter, and this, viewed against certain other facts, tipped the scale in favor of a "rigid" rating. On the face of it, counting Bender dots may be just as good (or poor) a sample of the compulsive-rigidity domain as straightening desk blotters is.

Suppose, to extend our example, we have four tests on the "predictor" side, over against the psychiatrist's "criterion," and find generally positive correlations among the five variables. Surely it is artificial and arbitrary to impose the "test-should-predict-criterion" pattern on such data. The psychiatrist samples verbal content, expressive pattern, voice, posture, etc. The psychologist samples verbal content, perception, expressive pattern, etc. Our proper conclusion is that, from this evidence, the four tests and the psychiatrist all assess some common factor.

The asymmetry between the "test" and the so-designated "criterion" arises only because the terminology of predictive validity has become a commonplace in test analysis. In this study where a construct is the central concern, any distinction between the merit of the test and criterion variables would be justified only if it had already been shown that the psychiatrist's theory and operations were excellent measures of the attribute.

### *Inadequacy of Validation in Terms of Specific Criteria*

The proposal to validate constructural interpretations of tests runs counter to suggestions of some others. Spiker and McCandless (57) favor an operational approach. Validation is replaced by compiling statements as to

how strongly the test predicts other observed variables of interest. To avoid requiring that each new variable be investigated completely by itself, they allow two variables to collapse into one whenever the properties of the operationally defined measures are the same: "If a new test is demonstrated to predict the scores on an older, well-established test, then an evaluation of the predictive power of the older test may be used for the new one." But accurate inferences are possible only if the two tests correlate so highly that there is negligible reliable variance in either test, independent of the other. Where the correspondence is less close, one must either retain all the separate variables operationally defined or embark on construct validation.

The practical user of tests must rely on constructs of some generality to make predictions about new situations. Test *X* could be used to predict palmar sweating in the face of failure without invoking any construct, but a counselor is more likely to be asked to forecast behavior in diverse or even unique situations for which the correlation of test *X* is unknown. Significant predictions rely on knowledge accumulated around the generalized construct of anxiety. The *Technical Recommendations* state:

It is ordinarily necessary to evaluate construct validity by integrating evidence from many different sources. The problem of construct validation becomes especially acute in the clinical field since for many of the constructs dealt with it is not a question of finding an imperfect criterion but of finding any criterion at all. The psychologist interested in construct validity for clinical devices is concerned with making an estimate of a hypothetical internal process, factor, system, structure, or state and cannot expect to find a clear unitary behavioral criterion. An attempt to identify any one criterion measure or any composite as *the* criterion aimed at is, however, usually unwarranted (59, p. 14-15).

This appears to conflict with arguments for specific criteria prominent at places in the testing literature. Thus Anastasi (2) makes many statements of the latter character: "It is only as a measure of a specifically defined criterion that a test can be objectively validated at all. . . . To claim that a test measures anything over and above its criterion is pure speculation" (p. 67). Yet elsewhere this article supports construct validation. Tests can be profitably interpreted if we "know the relationships between the tested behavior . . . and other behavior samples, none of these behavior samples necessarily occupying the preeminent position of a criterion" (p. 75). Factor analysis with several partial criteria might be used to study whether a test measures a postulated "general learning ability." If the data demonstrate specificity of ability instead, such specificity is "useful in its own right in advancing our knowledge of behavior; it should not be construed as a weakness of the tests" (p. 75).

We depart from Anastasi at two points. She writes, "The validity of a



psychological test should not be confused with an analysis of the factors which determine the behavior under consideration." We, however, regard such analysis as a most important type of validation. Second, she refers to "the will-o'-the-wisp of psychological processes which are distinct from performance" (2, p. 77). While we agree that psychological processes are elusive, we are sympathetic to attempts to formulate and clarify constructs which are evidenced by performance but distinct from it. Surely an inductive inference based on a pattern of correlations cannot be dismissed as "pure speculation."

#### SPECIFIC CRITERIA USED TEMPORARILY: THE "BOOTSTRAPS" EFFECT

Even when a test is constructed on the basis of a specific criterion, it may ultimately be judged to have greater construct validity than the criterion. We start with a vague concept which we associate with certain observations. We then discover empirically that these observations covary with some other observation which possesses greater reliability or is more intimately correlated with relevant experimental changes than is the original measure, or both. For example, the notion of temperature arises because some objects feel hotter to the touch than others. The expansion of a mercury column does not have face validity as an index of hotness. But it turns out that (a) there is a statistical relation between expansion and sensed temperature; (b) observers employ the mercury method with good interobserver agreement; (c) the regularity of observed relations is increased by using the thermometer (e.g., melting points of samples of the same materials vary little on the thermometer; we obtain nearly linear relations between mercury measures and pressure of a gas). Finally, (d) a theoretical structure involving unobservable microevents—the kinetic theory—is worked out which explains the relation of mercury expansion to heat. This whole process of conceptual enrichment begins with what in retrospect we see as an extremely fallible "criterion"—the human temperature sense. That original criterion has now been relegated to a peripheral position. We have lifted ourselves by our bootstraps, but in a legitimate and fruitful way.

Similarly, the Binet scale was first valued because children's scores tended to agree with judgments by schoolteachers. If it had not shown this agreement, it would have been discarded along with reaction time and the other measures of ability previously tried. Teacher judgments once constituted the criterion against which the individual intelligence test was validated. But if today a child's IQ is 135 and three of his teachers complain about how stupid he is, we do not conclude that the test has failed. Quite

to the contrary, if no error in test procedure can be argued, we treat the test score as a valid statement about an important quality, and define our task as that of finding out what other variables—personality, study skills, etc.—modify achievement or distort teacher judgment.

## *Experimentation to Investigate Construct Validity*

### VALIDATION PROCEDURES

We can use many methods in construct validation. Attention should particularly be drawn to Macfarlane's survey of these methods as they apply to projective devices (41).

*Group differences.* If our understanding of a construct leads us to expect two groups to differ on the test, this expectation may be tested directly. Thus Thurstone and Chave validated the Scale for Measuring Attitude Toward the Church by showing score differences between church members and nonchurchgoers. Churchgoing is not *the* criterion of attitude, for the purpose of the test is to measure something other than the crude sociological fact of church attendance; on the other hand, failure to find a difference would have seriously challenged the test.

Only coarse correspondence between test and group designation is expected. Too great a correspondence between the two would indicate that the test is to some degree invalid, because members of the groups are expected to overlap on the test. Intelligence test items are selected initially on the basis of a correspondence to age, but an item that correlates .95 with age in an elementary school sample would surely be suspect.

*Correlation matrices and factor analysis.* If two tests are presumed to measure the same construct, a correlation between them is predicted. (An exception is noted where some second attribute has positive loading in the first test and negative loading in the second test; then a low correlation is expected. This is a testable interpretation provided an external measure of either the first or the second variable exists.) If the obtained correlation departs from the expectation, however, there is no way to know whether the fault lies in test A, test B, or the formulation of the construct. A matrix of intercorrelations often points out profitable ways of dividing the construct into more meaningful parts, factor analysis being a useful computational method in such studies.

Guilford (26) has discussed the place of factor analysis in construct validation. His statements may be extracted as follows:



"The personnel psychologist wishes to know 'why his tests are valid.' He can place tests and practical criteria in a matrix and factor it to identify 'real dimensions of human personality.' A factorial description is exact and stable; it is economical in explanation; it leads to the creation of pure tests which can be combined to predict complex behaviors." It is clear that factors here function as constructs. Eysenck, in his "criterion analysis" (18), goes farther than Guilford, and shows that factoring can be used explicitly to test hypotheses about constructs.

Factors may or may not be weighted with surplus meaning. Certainly when they are regarded as "real dimensions" a great deal of surplus meaning is implied, and the interpreter must shoulder a substantial burden of proof. The alternative view is to regard factors as defining a working reference frame, located in a convenient manner in the "space" defined by all behaviors of a given type. Which set of factors from a given matrix is "most useful" will depend partly on predilections, but in essence the best construct is the one around which we can build the greatest number of inferences, in the most direct fashion.

*Studies of internal structure.* For many constructs, evidence of homogeneity within the test is relevant in judging validity. If a trait such as *dominance* is hypothesized, and the items inquire about behaviors subsumed under this label, then the hypothesis appears to require that these items be generally intercorrelated. Even low correlations, if consistent, would support the argument that people may be fruitfully described in terms of a generalized tendency to dominate or not dominate. The general quality would have power to predict behavior in a variety of situations represented by the specific items. Item-test correlations and certain reliability formulas describe internal consistency.

It is unwise to list uninterpreted data of this sort under the heading "validity" in test manuals, as some authors have done. High internal consistency may *lower* validity. Only if the underlying theory of the trait being measured calls for high item intercorrelations do the correlations support construct validity. Negative item-test correlations may support construct validity, provided that the items with negative correlations are believed irrelevant to the postulated construct and serve as suppressor variables (31, pp. 431-436; 44).

Study of distinctive subgroups of items within a test may set an upper limit to construct validity by showing that irrelevant elements influence scores. Thus a study of the PMA space tests shows that variance can be partially accounted for by a response set, tendency to mark many figures as similar (12). An internal factor analysis of the PEA Interpretation of Data Test shows that in addition to measuring reasoning skills, the test score is

strongly influenced by a tendency to say "probably true" rather than "certainly true," regardless of item content (17). On the other hand, a study of item groupings in the DAT Mechanical Comprehension Test permitted rejection of the hypothesis that knowledge about specific topics such as gears made a substantial contribution to scores (13).

*Studies of changeover occasions.* The stability of test scores ("retest reliability," Cattell's "N-technique") may be relevant to construct validation. Whether a high degree of stability is encouraging or discouraging for the proposed interpretation depends upon the theory defining the construct.

More powerful than the retest after uncontrolled intervening experiences is the retest with experimental intervention. If a transient influence swings test scores over a wide range, there are definite limits on the extent to which a test result can be interpreted as reflecting the typical behavior of the individual. These are examples of experiments which have indicated upper limits to test validity: studies of differences associated with the examiner in projective testing, of change of score under alternative directions ("tell the truth" vs. "make yourself look good to an employer"), and of coachability of mental tests. We may recall Gulliksen's distinction (27): When the coaching is of a sort that improves the pupil's intellectual functioning in school, the test which is affected by the coaching has validity as a measure of intellectual functioning; if the coaching improves test-taking but not school performance, the test which responds to the coaching has poor validity as a measure of this construct.

Sometimes, where differences between individuals are difficult to assess by any means other than the test, the experimenter validates by determining whether the test can detect induced intra-individual differences. One might hypothesize that the Zeigarnik effect is a measure of ego involvement, i.e., that with ego involvement there is more recall of incomplete tasks. To support such an interpretation, the investigator will try to induce ego involvement on some task by appropriate directions and compare subjects' recall with their recall for tasks where there was a contrary induction. Sometimes the intervention is drastic. Porteus finds (53) that brain-operated patients show disruption of performance on his maze, but do not show impaired performance on conventional verbal tests and argues therefrom that his test is a better measure of planfulness.

*Studies of process.* One of the best ways of determining informally what accounts for variability on a test is the observation of the person's process of performance. If it is supposed, for example, that a test measures mathematical competence, and yet observation of students' errors shows that erroneous reading of the question is common, the implications of a low score



are altered. Lucas in this way showed that the Navy Relative Movement Test, an aptitude test, actually involved two different abilities: spatial visualization and mathematical reasoning (39).

Mathematical analysis of scoring procedures may provide important negative evidence on construct validity. A recent analysis of "empathy" tests is perhaps worth citing (14). "Empathy" has been operationally defined in many studies by the ability of a judge to predict what responses will be given on some questionnaire by a subject he has observed briefly. A mathematical argument has shown, however, that the scores depend on several attributes of the judge which enter into his perception of *any* individual, and that they therefore cannot be interpreted as evidence of his ability to interpret cues offered by particular others, or his intuition.

#### THE NUMERICAL ESTIMATE OF CONSTRUCT VALIDITY

There is an understandable tendency to seek a "construct validity coefficient." A numerical statement of the degree of construct validity would be a statement of the proportion of the test score variance that is attributable to the construct variable. This numerical estimate can sometimes be arrived at by a factor analysis, but since present methods of factor analysis are based on linear relations, more general methods will ultimately be needed to deal with many quantitative problems of construct validation.

Rarely will it be possible to estimate definite "construct saturations," because no factor corresponding closely to the construct will be available. One can only hope to set upper and lower bounds to the "loading." If "creativity" is defined as something independent of knowledge, then a correlation of .40 between a presumed test of creativity and a test of arithmetic knowledge would indicate that at least 16 per cent of the reliable test variance is irrelevant to creativity as defined. Laboratory performance on problems such as Maier's "hatrack" would scarcely be an ideal measure of creativity, but it would be somewhat relevant. If its correlation with the test is .60, this permits a tentative estimate of 36 per cent as a lower bound. (The estimate is tentative because the test might overlap with the irrelevant portion of the laboratory measure.) The saturation seems to lie between 36 and 84 per cent; a cumulation of studies would provide better limits.

It should be particularly noted that rejecting the null hypothesis does not finish the job of construct validation (35, p. 284). The problem is not to conclude that the test "is valid" for measuring the construct variable. The task is to state as definitely as possible the degree of validity the test is presumed to have.

## *The Logic of Construct Validation*

Construct validation takes place when an investigator believes that his instrument reflects a particular construct, to which are attached certain meanings. The proposed interpretation generates specific testable hypotheses, which are a means of confirming or disconfirming the claim. The philosophy of science which we believe does most justice to actual scientific practice will now be briefly and dogmatically set forth. Readers interested in further study of the philosophical underpinning are referred to the works by Braithwaite (6, especially Chapter III), Carnap (7; 8, pp. 56-69), Pap (51), Sellars (55, 56), Feigl (19, 20), Beck (4), Kneale (37, pp. 92-110), Hempel (29; 30, Sec. 7).

### THE NOMOLOGICAL NET

The fundamental principles are these:

1. Scientifically speaking, to "make clear what something *is*" means to set forth the laws in which it occurs. We shall refer to the interlocking system of laws which constitute a theory as a *nomological network*.

2. The laws in a nomological network may relate (a) observable properties or quantities to each other; or (b) theoretical constructs to observables; or (c) different theoretical constructs to one another. These "laws" may be statistical or deterministic.

3. A necessary condition for a construct to be scientifically admissible is that it occur in a nomological net, at least *some* of whose laws involve observables. Admissible constructs may be remote from observation, i.e., a long derivation may intervene between the nomologicals which implicitly define the construct, and the (derived) nomologicals of type *a*. These latter propositions permit predictions about events. The construct is not "reduced" to the observations, but only combined with other constructs in the net to make predictions about observables.

4. "Learning more about" a theoretical construct is a matter of elaborating the nomological network in which it occurs, or of increasing the definiteness of the components. At least in the early history of a construct the network will be limited, and the construct will as yet have few connections.

5. An enrichment of the net such as adding a construct or a relation to theory is justified if it generates nomologicals that are confirmed by observation or if it reduces the number of nomologicals required to predict the same observations. When observations will not fit into the network as it



stands, the scientist has a certain freedom in selecting where to modify the network. That is, there may be alternative constructs or ways of organizing the net which for the time being are equally defensible.

6. We can say that "operations" which are qualitatively very different "overlap" or "measure the same thing" if their positions in the nomological net tie them to the same construct variable. Our confidence in this identification depends upon the amount of inductive support we have for the regions of the net involved. It is not necessary that a direct observational comparison of the two operations be made—we may be content with an intranetwork proof indicating that the two operations yield estimates of the same network-defined quantity. Thus, physicists are content to speak of the "temperature" of the sun and the "temperature" of a gas at room temperature even though the test operations are nonoverlapping because this identification makes theoretical sense.

With these statements of scientific methodology in mind, we return to the specific problem of construct validity as applied to psychological tests. The preceding guide rules should reassure the "toughminded," who fear that allowing construct validation opens the door to nonconfirmable test claims. *The answer is that unless the network makes contact with observations, and exhibits explicit, public steps of inference, construct validation cannot be claimed.* An admissible psychological construct must be behavior-relevant (59, p. 15). For most tests intended to measure constructs, adequate criteria do not exist. This being the case, many such tests have been left unvalidated, or a finespun network of rationalizations has been offered as if it were validation. Rationalization is not construct validation. One who claims that his test reflects a construct cannot maintain his claim in the face of recurrent negative results because these results show that his construct is too loosely defined to yield verifiable inferences.

A rigorous (though perhaps probabilistic) chain of inference is required to establish a test as a measure of a construct. To validate a claim that a test measures a construct, a nomological net surrounding the concept must exist. When a construct is fairly new, there may be few specifiable associations by which to pin down the concept. As research proceeds, the construct sends out roots in many directions, which attach it to more and more facts or other constructs. Thus the electron has more accepted properties than the neutrino; *numerical ability* has more than *the second space factor*.

"Acceptance," which was critical in criterion-oriented and content validities, has now appeared in construct validity. Unless substantially the same nomological net is accepted by the several users of the construct, public validation is impossible. If A uses *aggressiveness* to mean overt assault on others, and B's usage includes repressed hostile reactions, evidence which

convinces B that a test measures *aggressiveness* convinces A that the test does not. Hence, the investigator who proposes to establish a test as a measure of a construct must specify his network or theory sufficiently clearly that others can accept or reject it (cf. 41, p. 406). A consumer of the test who rejects the author's theory cannot accept the author's validation. He must validate the test for himself, if he wishes to show that it represents the construct as *he* defines it.

Two general qualifications are in order with reference to the methodological principles 1-6 set forth at the beginning of this section. Both of them concern the amount of "theory," in any high-level sense of that word, which enters into a construct-defining network of laws or lawlike statements. We do not wish to convey the impression that one always has a very elaborate theoretical network, rich in hypothetical processes or entities.

*Constructs as inductive summaries.* In the early stages of development of a construct or even at more advanced stages when our orientation is thoroughly practical, little or no theory in the usual sense of the word need be involved. In the extreme case the hypothesized laws are formulated entirely in terms of descriptive (observational) dimensions although not all of the relevant observations have actually been made.

The hypothesized network "goes beyond the data" only in the limited sense that it purports to *characterize* the behavior facets which belong to an observable but as yet only partially sampled cluster; hence, it generates predictions about hitherto unsampled regions of the phenotypic space. Even though no unobservables or high-order theoretical constructs are introduced, an element of inductive extrapolation appears in the claim that a cluster including some elements not-yet-observed has been identified. Since, as in any sorting or abstracting task involving a finite set of complex elements, several nonequivalent bases of categorization are available, the investigator may choose a hypothesis which generates erroneous predictions. The failure of a supposed, hitherto untried, member of the cluster to behave in the manner said to be characteristic of the group, or the finding that a nonmember of the postulated cluster does behave in this manner, may modify greatly our tentative construct.

For example, one might build an intelligence test on the basis of his background notions of "intellect," including vocabulary, arithmetic calculation, general information, similarities, two-point threshold, reaction time, and line bisection as subtests. The first four of these correlate, and he extracts a huge first factor. This becomes a second approximation of the intelligence construct, described by its pattern of loadings on the four tests. The other three tests have negligible loading on any common factor. On this evidence the investigator reinterprets intelligence as "manipulation of



words." Subsequently it is discovered that test-stupid people are rated as unable to express their ideas, are easily taken in by fallacious arguments, and misread complex directions. These data support the "linguistic" definition of intelligence and the test's claim of validity for that construct. But then a block design test with pantomime instructions is found to be strongly saturated with the first factor. Immediately the purely "linguistic" interpretation of Factor I becomes suspect. This finding, taken together with our initial acceptance of the others as relevant to the background concept of intelligence, forces us to reinterpret the concept once again.

If we simply *list* the tests or traits which have been shown to be saturated with the "factor" or which belong to the cluster, no construct is employed. As soon as we even *summarize the properties* of this group of indicators—we are already making some guesses. Intentional characterization of a domain is hazardous since it selects (abstracts) properties and implies that new tests sharing those properties will behave as do the known tests in the cluster, and that tests not sharing them will not.

The difficulties in merely "characterizing the surface cluster" are strikingly exhibited by the use of certain special and extreme groups for purposes of construct validation. The  $P_d$  scale of MMPI was originally derived and cross-validated upon hospitalized patients diagnosed "Psychopathic personality, asocial and amoral type" (42). Further research shows the scale to have a limited degree of predictive and concurrent validity for "delinquency" more broadly defined (5, 28). Several studies show associations between  $P_d$  and very special "criterion" groups which it would be ludicrous to identify as "the criterion" in the traditional sense. If one lists these heterogeneous groups and tries to characterize them intentionally, he faces enormous conceptual difficulties. For example, a recent survey of hunting accidents in Minnesota showed that hunters who had "carelessly" shot someone were significantly elevated on  $P_d$  when compared with other hunters (48). This is in line with one's theoretical expectations; when you ask MMPI "experts" to predict for such a group they invariably predict  $P_d$  or  $M_d$  or both. The finding seems therefore to lend some slight support to the construct validity of the  $P_d$  scale. But of course it would be nonsense to *define* the  $P_d$  component "operationally" in terms of, say, accident proneness. We might try to subsume the original phenotype and the hunting-accident proneness under some broader category, such as "Disposition to violate society's rules, whether legal, moral, or just *sensible*." But now we have ceased to have a neat operational criterion, and are using instead a rather vague and wide-range class. Besides, there is worse to come. We want the class specification to cover a group trend that (nondelinquent) high school students judged by their peer group as least "responsible" score over a full sigma higher on  $P_d$

than those judged most "responsible" (23, p. 75). Most of the behaviors contributing to such sociometric choices fall well within the range of socially permissible action; the proffered criterion specification is still too restrictive. Again, any clinician familiar with MMPI lore would predict an elevated  $P_d$  on a sample of (nondelinquent) professional actors. Chyatte's confirmation of this prediction (10) tends to support *both*: (a) the theory sketch of "what the  $P_d$  factor is, psychologically"; and (b) the claim of the  $P_d$  scale to construct validity for this hypothetical factor. Let the reader try his hand at writing a brief phenotypic criterion specification that will cover both trigger-happy hunters and Broadway actors! And if he should be ingenious enough to achieve this, does his definition also encompass Hovey's report that high  $P_d$  predicts the judgments "not shy" and "unafraid of mental patients" made upon nurses by their supervisors (32, p. 143)? And then we have Gough's report that *low*  $P_d$  is associated with ratings as "good-natured" (24, p. 40), and Roessell's data showing that high  $P_d$  is predictive of "dropping out of high school" (54). The point is that all seven of these "criterion" dispositions would be readily guessed by any clinician having even superficial familiarity with MMPI interpretation; but to mediate these inferences explicitly requires quite a few hypotheses about dynamics, constituting an admittedly sketchy (but far from vacuous) network defining the genotype *psychopathic deviate*.

*Vagueness of present psychological laws.* This line of thought leads directly to our second important qualification upon the network schema. The idealized picture is one of a tidy set of postulates which jointly entail the desired theorems; since some of the theorems are coordinated to the observation base, the system constitutes an implicit definition of the theoretical primitives and gives them an indirect empirical meaning. In practice, of course, even the most advanced physical sciences only approximate this ideal. Questions of "categoricalness" and the like, such as logicians raise about pure calculi, are hardly even statable for empirical networks. (What, for example, would be the desiderata of a "well-formed formula" in molar behavior theory?) Psychology works with crude, half-explicit formulations. We do not worry about such advanced formal questions as "whether all molar-behavior statements are decidable by appeal to the postulates" because we know that no existing theoretical network suffices to predict even the *known* descriptive laws. Nevertheless, the sketch of a network is there; if it were not, we would not be saying *anything* intelligible about our constructs. We do not have the rigorous implicit definitions of formal calculi (which still, be it noted, usually permit of a multiplicity of interpretations). Yet the vague, avowedly incomplete network still gives the constructs what-



ever meaning they do have. When the network is very incomplete, having many strands missing entirely and some constructs tied in only by tenuous threads, then the "implicit definition" of these constructs is disturbingly loose; one might say that the meaning of the constructs is underdetermined. *Since the meaning of theoretical constructs is set forth by stating the laws in which they occur, our incomplete knowledge of the laws of nature produces a vagueness in our constructs* (See Hempel [30]; Kaplan [34]; Pap [51]). We will be able to say "what anxiety is" when we know all of the laws involving it; meanwhile, since we are in the process of discovering these laws, we do not yet know precisely what anxiety is.

### *Conclusions Regarding the Network After Experimentation*

The proposition that  $x$  per cent of test variance is accounted for by the construct is inserted into the accepted network. The network then generates a testable prediction about the relation of the test scores to certain other variables, and the investigator gathers data. If prediction and result are in harmony, he can retain his belief that the test measures the construct. The construct is at best adopted, never demonstrated to be "correct."

We do not first "prove" the theory, and then validate the test, nor conversely. In any probable inductive type of inference from a pattern of observations, we examine the relation between the total network of theory and observations. The system involves propositions relating test to construct, construct to other constructs, and finally relating some of these constructs to observables. In ongoing research the chain of inference is very complicated. Kelly and Fiske (36, p. 124) give a complex diagram showing the numerous inferences required in validating a prediction from assessment techniques, where theories about the criterion situation are as integral a part of the prediction as are the test data. A predicted empirical relationship permits us to test all the propositions leading to that prediction. Traditionally the proposition claiming to interpret the test has been set apart as the hypothesis being tested, but actually the evidence is significant for all parts of the chain. If the prediction is not confirmed, any link in the chain may be wrong.

A theoretical network can be divided into subtheories used in making particular predictions. All the events successfully predicted through a subtheory are of course evidence in favor of that theory. Such a subtheory may be so well confirmed by voluminous and diverse evidence that we can rea-

sonably view a particular experiment as relevant only to the test's validity. If the theory, combined with a proposed test interpretation, mispredicts in this case, it is the latter which must be abandoned. On the other hand, the accumulated evidence for a test's construct validity may be so strong that an instance of misprediction will force us to modify the subtheory employing the construct rather than deny the claim that the test measures the construct.

Most cases in psychology today lie somewhere between these extremes. Thus, suppose we fail to find a greater incidence of "homosexual signs" in the Rorschach records of paranoid patients. Which is more strongly disconfirmed—the Rorschach signs or the orthodox theory of paranoia? The negative finding shows the bridge between the two to be undependable, but this is all we can say. The bridge cannot be used unless one end is placed on solid ground. The investigator must decide which end it is best to relocate.

Numerous successful predictions dealing with phenotypically diverse "criteria" give greater weight to the claim of construct validity than do fewer predictions, or predictions involving very similar behaviors. In arriving at diverse predictions, the hypothesis of test validity is connected each time to a subnetwork largely independent of the portion previously used. Success of these derivations testifies to the inductive power of the test-validity statement, and renders it unlikely that an equally effective alternative can be offered.

#### IMPLICATIONS OF NEGATIVE EVIDENCE

The investigator whose prediction and data are discordant must make strategic decisions. His result can be interpreted in three ways:

1. The test does not measure the construct variable.
2. The theoretical network which generated the hypothesis is incorrect.
3. The experimental design failed to test the hypothesis properly. (Strictly speaking this may be analyzed as a special case of 2, but in practice the distinction is worth making.)

*For further research.* If a specific fault of procedure makes the third a reasonable possibility, his proper response is to perform an adequate study, meanwhile making no report. When faced with the other two alternatives, he may decide that his test does not measure the construct adequately. Following that decision, he will perhaps prepare and validate a new test. Any rescoring or new interpretative procedure for the original instrument, like a new test, requires validation *by means of a fresh body of data*.

The investigator may regard Interpretation 2 as more likely to lead to eventual advances. It is legitimate for the investigator to call the network defining the construct into question, if he has confidence in the test. Should



the investigator decide that some step in the network is unsound, he may be able to invent an alternative network. Perhaps he modifies the network by splitting a concept into two or more portions, e.g., by designating types of *anxiety*, or perhaps he specifies added conditions under which a generalization holds. When an investigator modifies the theory in such a manner, he is now required to *gather a fresh body of data* to test the altered hypotheses. This step should normally precede publication of the modified theory. If the new data are consistent with the modified network, he is free from the fear that his nomologicals were gerrymandered to fit the peculiarities of his first sample of observations. He can now trust his test to some extent, because his test results behave as predicted.

The choice among alternatives, like any strategic decision, is a gamble as to which course of action is the best investment of effort. Is it wise to modify the theory? That depends on how well the system is confirmed by prior data, and how well the modifications fit available observations. Is it worth while to modify the test in the hope that it will fit the construct? That depends on how much evidence there is—apart from this abortive experiment—to support the hope, and also on how much it is worth to the investigator's ego to salvage the test. The choice among alternatives is a matter of research planning.

*For practical use of the test.* The consumer can accept a test as a measure of a construct only when there is a strong positive fit between predictions and subsequent data. When the evidence from a proper investigation of a published test is essentially negative, it should be reported as a stop sign to discourage use of the test pending a reconciliation of test and construct, or final abandonment of the test. If the test has not been published, it should be restricted to research use until some degree of validity is established (1). The consumer can await the results of the investigator's gamble with confidence that proper application of the scientific method will ultimately tell whether the test has value. Until the evidence is in, he has no justification for employing the test as a basis for terminal decisions. The test may serve, at best, only as a source of suggestions about individuals to be confirmed by other evidence (15, 47).

There are two perspectives in test validation. From the viewpoint of the psychological practitioner, the burden of proof is on the test. A test should not be used to measure a trait until its proponent establishes that predictions made from such measures are consistent with the best available theory of the trait. In the view of the test developer, however, both the test and the theory are under scrutiny. He is free to say *to himself privately*, "If my test disagrees with the theory, so much the worse for the theory." This way lies delusion, unless he continues his research using a better theory.

## REPORTING OF POSITIVE RESULTS

The test developer who finds positive correspondence between his proposed interpretation and data is expected to report the basis for his validity claim. Defending a claim of construct validity is a major task, not to be satisfied by a discourse without data. The *Technical Recommendations* have little to say on reporting of construct validity. Indeed, the only detailed suggestions under that heading refer to correlations of the test with other measures, together with a cross reference to some other sections of the report. The two key principles, however, call for the most comprehensive type of reporting. The manual for any test "should report all available information which will assist the user in determining what psychological attributes account for variance in test scores" (59, p. 27). And, "The manual for a test which is used primarily to assess postulated attributes of the individual should outline the theory on which the test is based and organize whatever partial validity data there are to show in what way they support the theory" (59, p. 28). It is recognized, by a classification as "very desirable" rather than "essential," that the latter recommendation goes beyond present practice of test authors.

The proper goals in reporting construct validation are to make clear (a) what interpretation is proposed, (b) how adequately the writer believes this interpretation is substantiated, and (c) what evidence and reasoning lead him to this belief. Without *a* the construct validity of the test is of no use to the consumer. Without *b* the consumer must carry the entire burden of evaluating the test research. Without *c* the consumer or reviewer is being asked to take *a* and *b* on faith. The test manual cannot always present an exhaustive statement on these points, but it should summarize and indicate where complete statements may be found.

To specify the interpretation, the writer must state what construct he has in mind, and what meaning he gives to that construct. For a construct which has a short history and has built up few connotations, it will be fairly easy to indicate the presumed properties of the construct, i.e., the nomologicals in which it appears. For a construct with a longer history, a summary of properties and references to previous theoretical discussions may be appropriate. It is especially critical to distinguish proposed interpretations from other meanings previously given the same construct. The validator faces no small task; he must somehow communicate a theory to his reader.

To evaluate his evidence calls for a statement like the conclusions from a program of research, noting what is well substantiated and what alternative interpretations have been considered and rejected. The writer must note what portions of his proposed interpretation are speculations, extrapo-



lations, or conclusions from insufficient data. The author has an ethical responsibility to prevent unsubstantiated interpretations from appearing as truths. A claim is unsubstantiated unless the evidence for the claim is public, so that other scientists may review the evidence, criticize the conclusions, and offer alternative interpretations.

The report of evidence in a test manual must be as nearly complete as any research report, except where adequate public reports can be cited. Reference to something "observed by the writer in many clinical cases" is worthless as evidence. Full case reports, on the other hand, may be a valuable source of evidence so long as these cases are representative and negative instances receive due attention. The report of evidence must be interpreted with reference to the theoretical network in such a manner that the reader sees why the author regards a particular correlation or experiment as confirming (or throwing doubt upon) the proposed interpretation. Evidence collected by others must be taken fairly into account.

### *Validation of a Complex Test "As a Whole"*

Special questions must be considered when we are investigating the validity of a test which is aimed to provide information about several constructs. In one sense, it is naïve to inquire "Is this test valid?" One does not validate a test, but only a principle for making inferences. If a test yields many different types of inferences, some of them can be valid and others invalid (cf. Technical Recommendation C2: "The manual should report the validity of each type of inference for which a test is recommended"). From this point of view, every topic sentence in the typical book on Rorschach interpretation presents a hypothesis requiring validation, and one should validate inferences about each aspect of the personality separately and in turn, just as he would want information on the validity (concurrent or predictive) for each scale of MMPI.

There is, however, another defensible point of view. If a test is purely empirical, based strictly on observed connections between response to an item and some criterion, then of course the validity of one scoring key for the test does not make validation for its other scoring keys any less necessary. But a test may be developed on the basis of a theory which in itself provides a linkage between the various keys and the various criteria. Thus, while Strong's Vocational Interest Blank is developed empirically, it also rests on a "theory" that a youth can be expected to be satisfied in an occupation if he has interests common to men now happy in the occupation. When Strong finds that those with high engineering interest scores in college are

preponderantly in engineering careers nineteen years later, he has partly validated the proposed use of the engineer score (predictive validity). Since the evidence is consistent with the theory on which all the test keys were built, this evidence alone increases the presumption that the *other* keys have predictive validity. How strong is this presumption? Not very, from the viewpoint of the traditional skepticism of science. Engineering interests may stabilize early, while interests in art or management or social work are still unstable. A claim cannot be made that the whole Strong approach is valid just because one score shows predictive validity. But if thirty interest scores were investigated longitudinally and all of them showed the type of validity predicted by Strong's theory, we would indeed be caviling to say that this evidence gives no confidence in the long-range validity of the thirty-first score.

Confidence in a theory is increased as more relevant evidence confirms it, but it is always possible that tomorrow's investigation will render the theory obsolete. The *Technical Recommendations* suggest a rule of reason, and ask for evidence for each *type* of inference for which a test is recommended. It is stated that no test developer can present predictive validities for all possible criteria; similarly, no developer can run all possible experimental tests of his proposed interpretation. But the recommendation is more subtle than advice that a lot of validation is better than a little.

Consider the Rorschach test. It is used for many inferences, made by means of nomological networks at several levels. At a low level are the simple unrationalized correspondences presumed to exist between certain signs and psychiatric diagnoses. Validating such a sign does nothing to substantiate Rorschach theory. For other Rorschach formulas an explicit *a priori* rationale exists (for instance, high *F*% interpreted as implying rigid control of impulses). Each time such a sign shows correspondence with criteria, its rationale is supported just a little. At a still higher level of abstraction, a considerable body of theory surrounds the general area of *outer control*, interlacing many different constructs. As evidence cumulates, one should be able to decide what specific inference-making chains within this system can be depended upon. One should also be able to conclude—or deny—that so much of the system has stood up under test that one has some confidence in even the untested lines in the network.

In addition to relatively delimited nomological networks surrounding *control* or *aspiration*, the Rorschach interpreter usually has an overriding theory of the test as a whole. This may be a psychoanalytic theory, a theory of perception and set, or a theory stated in terms of learned habit patterns. Whatever the theory of the interpreter, whenever he validates an inference from the system, he obtains some reason for added confidence in his overriding system. His total theory is not tested, however, by experiments dealing



with only one limited set of constructs. The test developer must investigate far-separated, independent sections of the network. The more diversified the predictions the system is required to make, the greater confidence we can have that only minor parts of the system will later prove faulty. Here we begin to glimpse a logic to defend the judgment that the test and its whole interpretative system is valid at some level of confidence.

There are enthusiasts who would conclude from the foregoing paragraphs that since there is some evidence of correct, diverse predictions made from the Rorschach, the test as a whole can now be accepted as validated. This conclusion overlooks the negative evidence. Just one finding contrary to expectation, based on sound research, is sufficient to wash a whole theoretical structure away. Perhaps the remains can be salvaged to form a new structure. But this structure now must be exposed to fresh risks, and sound negative evidence will destroy it in turn. There is sufficient negative evidence to prevent acceptance of the Rorschach and its accompanying interpretative structures as a whole. So long as any aspects of the overriding theory stated for the test have been disconfirmed, this structure must be rebuilt.

Talk of areas and structures may seem not to recognize those who would interpret the personality "globally." They may argue that a test is best validated in matching studies. Without going into detailed questions of matching methodology, we can ask whether such a study validates the nomological network "as a whole." The judge does employ some network in arriving at his conception of his subject, integrating specific inferences from specific data. Matching studies, if successful, demonstrate only that each judge's interpretative theory has some validity, that it is not completely a fantasy. Very high consistency between judges is required to show that they are using the same network, and very high success in matching is required to show that the network is dependable.

If inference is less than perfectly dependable, we must know which aspects of the interpretative network are least dependable and which are most dependable. Thus, even if one has considerable confidence in a test "as a whole" because of frequent successful inferences, one still returns as an ultimate aim to the request of the technical recommendation for separate evidence on the validity of each type of inference to be made.

### *Recapitulation*

Construct validation was introduced in order to specify types of research required in developing tests for which the conventional views on validation are inappropriate. Personality tests, and some tests of ability, are interpreted in terms of attributes for which there is no adequate criterion. This paper

indicates what sort of evidence can substantiate such an interpretation, and how such evidence is to be interpreted. The following points made in the discussion are particularly significant.

1. A construct is defined implicitly by a network of associations or propositions in which it occurs. Constructs employed at different stages of research vary in definiteness.

2. Construct validation is possible only when some of the statements in the network lead to predicted relations among observables. While some observables may be regarded as "criteria," the construct validity of the criteria themselves is regarded as under investigation.

3. The network defining the construct, and the derivation leading to the predicted observation, must be reasonably explicit so that validating evidence may be properly interpreted.

4. Many types of evidence are relevant to construct validity, including content validity, interitem correlations, intertest correlations, test-"criterion" correlations, studies of stability over time, and stability under experimental intervention. High correlations and high stability may constitute either favorable or unfavorable evidence for the proposed interpretation, depending on the theory surrounding the construct.

5. When a predicted relation fails to occur, the fault may lie in the proposed interpretation of the test or in the network. Altering the network so that it can cope with the new observations is, in effect, redefining the construct. Any such new interpretation of the test must be validated by a fresh body of data before being advanced publicly. Great care is required to avoid substituting a posteriori rationalizations for proper validation.

6. Construct validity cannot generally be expressed in the form of a single simple coefficient. The data often permit one to establish upper and lower bounds for the proportion of test variance which can be attributed to the construct. The integration of diverse data into a proper interpretation cannot be an entirely quantitative process.

7. Constructs may vary in nature from those very close to "pure description" (involving little more than extrapolation of relations among observation-variables) to highly theoretical constructs involving hypothesized entities and processes, or making identifications with constructs of other sciences.

8. The investigation of a test's construct validity is not essentially different from the general scientific procedures for developing and confirming theories.

Without in the least *advocating* construct validity as preferable to the other three kinds (concurrent, predictive, content), we do believe it im-



perative that psychologists make a place for it in their methodological thinking, so that its rationale, its scientific legitimacy, and its dangers may become explicit and familiar. This would be preferable to the widespread current tendency to engage in what actually amounts to construct validation research and use of constructs in practical testing, while talking an "operational" methodology which, if adopted, would force research into a mold it does not fit.

## REFERENCES

1. AMERICAN PSYCHOLOGICAL ASSOCIATION. *Ethical standards of psychologists*. Washington, D.C.: American Psychological Association, Inc., 1953.
2. ANASTASI, ANNE. The concept of validity in the interpretation of test scores. *Educ. Psychol. Measmt.*, 1950, 10, 67-78.
3. BECHTOLDT, H. P. Selection. In S. S. Stevens (ed.), *Handbook of experimental psychology*. New York: Wiley, 1951. Pp. 1237-1267.
4. BECK, L. W. Constructions and inferred entities. *Phil. Sci.*, 1950, 17. Reprinted in H. Feigl and M. Brodbeck (eds.), *Readings in the philosophy of science*. New York: Appleton-Century-Crofts, 1953. Pp. 368-381.
5. BLAIR, W. R. N. A comparative study of disciplinary offenders and non-offenders in the Canadian Army. *Canad. J. Psychol.*, 1950, 4, 49-62.
6. BRAITHWAITE, R. B. *Scientific explanation*. Cambridge Univer. Press, 1953.
7. CARNAP, R. Empiricism, semantics, and ontology. *Rév. Int. de Phil.*, 1950, II, 20-40. Reprinted in P. P. Wiener (ed.), *Readings in philosophy of science*. New York: Scribner's, 1953. Pp. 509-521.
8. CARNAP, R. *Foundations of logic and mathematics*. *International encyclopedia of unified science*, I, No. 3. Pages 56-69 reprinted as "The interpretation of physics" in H. Feigl and M. Brodbeck (eds.), *Readings in the philosophy of science*. New York: Appleton-Century-Crofts, 1953. Pp. 309-318.
9. CHILD, I. L. Personality. *Annu. Rev. Psychol.*, 1954, 5, 149-171.
10. CHYATTE, C. Psychological characteristics of a group of professional actors. *Occupations*, 1949, 27, 245-250.
11. CRONBACH, L. J. *Essentials of psychological testing*. New York: Harper, 1949.
12. CRONBACH, L. J. Further evidence on response sets and test design. *Educ. Psychol. Measmt.*, 1950, 10, 3-31.
13. CRONBACH, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*, 1951, 16, 297-335.
14. CRONBACH, L. J. Processes affecting scores on "understanding of others" and "assumed similarity." *Psychol. Bull.*, 1955, 52, 177-193.
15. CRONBACH, L. J. The counselor's problems from the perspective of communication theory. In Vivian H. Hewer (ed.), *New perspectives in counseling*. Minneapolis: Univer. of Minnesota Press, 1955.

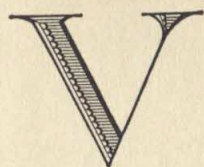
16. CURETON, E. E. Validity. In E. F. Lindquist (ed.), *Educational measurement*. Washington, D.C.: American Council on Education, 1950. Pp. 621-695.
17. DAMRIN, DORA E. A comparative study of information derived from a diagnostic problem-solving test by logical and factorial methods of scoring. Unpublished doctor's dissertation, Univer. of Illinois, 1952.
18. EYSENCK, H. J. Criterion analysis—an application of the hypothetico-deductive method in factor analysis. *Psychol. Rev.*, 1950, 57, 38-53.
19. FEIGL, H. Existential hypotheses. *Phil. Sci.*, 1950, 17, 35-62.
20. FEIGL, H. Confirmability and confirmation. *Rév. Int. de Phil.*, 1951, 5, 1-12. Reprinted in P. P. Wiener (ed.), *Readings in philosophy of science*. New York: Scribner's, 1953. Pp. 522-530.
21. GAYLORD, R. H. Conceptual consistency and criterion equivalence: a dual approach to criterion analysis. Unpublished manuscript (PRB Research Note No. 17). Copies obtainable from ASTIA-DSC, AD-21 440.
22. GOODENOUGH, FLORENCE L. *Mental testing*. New York: Rinehart, 1950.
23. GOUGH, H. G., MCCLOSKEY, H., & MEEHL, P. E. A personality scale for social responsibility. *J. Abnorm. Soc. Psychol.*, 1952, 47, 73-80.
24. GOUGH, H. G., MCKEE, M. G., & YANDELL, R. J. Adjective check list analyses of a number of selected psychometric and assessment variables. Unpublished manuscript. Berkeley: IPAR, 1953.
25. GUILFORD, J. P. New standards for test evaluation. *Educ. Psychol. Measmt.*, 1946, 6, 427-439.
26. GUILFORD, J. P. Factor analysis in a test-development program. *Psychol. Rev.*, 1948, 55, 79-94.
27. GULLIKSEN, H. Intrinsic validity. *Amer. Psychologist*, 1950, 5, 511-517.
28. HATHAWAY, S. R., & MONACHESI, E. D. *Analyzing and predicting juvenile delinquency with the MMPI*. Minneapolis: Univer. of Minnesota Press, 1953.
29. HEMPEL, C. G. Problems and changes in the empiricist criterion of meaning. *Rév. Int. de Phil.*, 1950, 4, 41-63. Reprinted in L. Linsky, *Semantics and the philosophy of language*. Urbana: Univer. of Illinois Press, 1952. Pp. 163-185.
30. HEMPEL, C. G. *Fundamentals of concept formation in empirical science*. Chicago: Univer. of Chicago Press, 1952.
31. HORST, P. The prediction of personal adjustment. *Soc. Sci. Res. Council Bull.*, 1941, No. 48.
32. HOVEY, H. B. MMPI profiles and personality characteristics. *J. Consult. Psychol.*, 1953, 17, 142-146.
33. JENKINS, J. G. Validity for what? *J. Consult. Psychol.*, 1946, 10, 93-98.
34. KAPLAN, A. Definition and specification of meaning. *J. Phil.*, 1946, 43, 281-288.
35. KELLY, E. L. Theory and techniques of assessment. *Annu. Rev. Psychol.*, 1954, 5, 281-311.
36. KELLY, E. L., & FISKE, D. W. *The prediction of performance in clinical psychology*. Ann Arbor: Univer. of Michigan Press, 1951.
37. KNEALE, W. *Probability and induction*. Oxford: Clarendon Press, 1949. Pages 92-110 reprinted as "Induction, explanation, and transcendent hypotheses"



- in H. Feigl and M. Brodbeck (eds.), *Readings in the philosophy of science*. New York: Appleton-Century-Crofts, 1953. Pp. 353-367.
38. LINDQUIST, E. F. *Educational measurement*. Washington, D.C.: American Council on Education, 1950.
  39. LUCAS, C. M. Analysis of the relative movement test by a method of individual interviews. *Bur. Naval Personnel Res. Rep.*, Contract Nonr-690 (00), NR 151-13, Educational Testing Service, March 1953.
  40. MACCORQUODALE, K., & MEEHL, P. E. On a distinction between hypothetical constructs and intervening variables. *Psychol. Rev.*, 1948, 55, 95-107.
  41. MACFARLANE, JEAN W. Problems of validation inherent in projective methods. *Amer. J. Orthopsychiat.*, 1942, 12, 405-410.
  42. MCKINLEY, J. C., & HATHAWAY, S. R. The MMPI: V. Hysteria, hypomania, and psychopathic deviate. *J. Appl. Psychol.*, 1944, 28, 153-174.
  43. MCKINLEY, J. C., HATHAWAY, S. R., & MEEHL, P. E. The MMPI: VI. The K scale. *J. Consult. Psychol.*, 1948, 12, 20-31.
  44. MEEHL, P. E. A simple algebraic development of Horst's suppressor variables. *Amer. J. Psychol.*, 1945, 58, 550-554.
  45. MEEHL, P. E. An investigation of a general normality or control factor in personality testing. *Psychol. Monogr.*, 1945, 59, No. 4 (whole no. 274).
  46. MEEHL, P. E. *Clinical vs. statistical prediction*. Minneapolis: Univer. of Minnesota Press, 1954.
  47. MEEHL, P. E., & ROSEN, A. Antecedent probability and the efficiency of psychometric signs, patterns or cutting scores. *Psychol. Bull.*, 1955, 52, 194-216.
  48. *Minnesota Hunter Casualty Study*. St. Paul: Jacob Schmidt Brewing Company, 1954.
  49. MOSIER, C. I. A critical examination of the concepts of face validity. *Educ. Psychol. Measmt.*, 1947, 7, 191-205.
  50. MOSIER, C. I. Problems and designs of cross-validation. *Educ. Psychol. Measmt.*, 1951, 11, 5-12.
  51. PAP, A. Reduction-sentences and open concepts. *Methods*, 1953, 5, 3-30.
  52. PEAK, HELEN. Problems of objective observation. In L. Festinger and D. Katz (eds.), *Research methods in the behavioral sciences*. New York: Dryden Press, 1953. Pp. 243-300.
  53. PORTEUS, S. D. *The Porteus maze test and intelligence*. Palo Alto: Pacific Books, 1950.
  54. ROESSEL, F. P. MMPI results for high school drop-outs and graduates. Unpublished doctor's dissertation, Univer. of Minnesota, 1954.
  55. SELLARS, W. S. Concepts as involving laws and inconceivable without them. *Phil. Sci.*, 1948, 15, 287-315.
  56. SELLARS, W. S. Some reflections on language games. *Phil. Sci.*, 1954, 21, 204-228.
  57. SPIKER, C. C., & MCCANDLESS, B. R. The concept of intelligence and the philosophy of science. *Psychol. Rev.*, 1954, 61, 255-267.
  58. Technical recommendations for psychological tests and diagnostic techniques: preliminary proposal. *Amer. Psychologist*, 1952, 7, 461-476.

59. Technical recommendations for psychological tests and diagnostic techniques. *Psychol. Bull. Supplement*, 1954, 51, 2, Part 2, 1-38.
60. THURSTONE, L. L. The criterion problem in personality research. *Psychometric Lab. Rep.*, No. 78. Chicago: Univer. of Chicago, 1952.

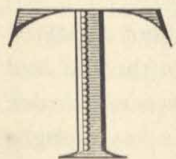




*Reliability*

Marguerite R. Hertz

# THE RELIABILITY OF THE RORSCHACH INK-BLOT TEST<sup>1</sup>

HE RORSCHACH Ink-blot Test<sup>2</sup> has been applied in many fields—abnormal psychology, psychiatry, mental tests, psychology of perception, heredity, childhood and adolescence, educational psychology and vocational guidance. The technique, the symptomatic value of the various categories and the interpretation of the test as a whole have been subject to extensive investigation. The method has been found of value in making personality studies of school children and adults, in studying racial differences, adolescent changes in personality, and the influence of environment and heredity on mental make-up, also in diagnosing general intelligence, types of intelligence and grades of mental deficiency. It has likewise been studied in connection with various other typological systems—constitutional types, eidetic types, graphological types and form-color types. It has been used to reveal imagination, inventiveness, fantasy living, poetic talent, and various aesthetic values, and to detect mental and moral defects, schizophrenic, neurotic, psy-

Reprinted from *J. Appl. Psychol.*, 1934, 18, 461-477, by permission of the American Psychological Association and the author.

1. Published through the courtesy of the Richman Fund.

2. For a description of the method, reference should be made to Rorschach (18), Rorschach and Oberholzer (19). Summaries of the method may be found in Behn-Eschenburg (4), Enke (5), Soukup (20), Müller (12), Oeser (14), Löpfe (10), Loosli-Usteri (9), Pfahler (15), Beck (2, 3), Vernon (22), and Hertz (8).



chotic traits and the like. Finally it has been used as a clinical instrument for making diagnosis and has been found of value in psychoanalysis.<sup>3</sup>

A review of the literature discloses a shocking scarcity of data on the reliability of the method. With few exceptions, it has not even been challenged. Despite extensive investigation and application little thought seems to have been given to the problem of whether or not it diagnoses personality consistently and reliably. Without doubt the results of the studies made with the method depend upon its reliability and must be interpreted with caution until that is established. Before the Rorschach test can be used as an instrument of measurement and diagnosis, it must show consistency of results when applied repeatedly to the same individuals under similar conditions.

### *The Problem of Reliability*

A number of factors make for inconsistency of results. They include 1) lack of standardization of procedure, 2) too limited a sampling of the individual's behavior, 3) unreliability of the sampling of the individual's behavior because of the variability of his performance, and 4) lack of objectivity in scoring.

It was observed in working with the Rorschach Test that there appeared to be great variability in the reactions of divers individuals to the test and that such wide and variable performance might not give an accurate index in the reactions of the same individual at different sittings. It was thought that such wide and variable performance might not give an accurate index of the individual's habitual behavior or personality. It was of greatest importance, therefore, to study the variability-performance factor to determine whether the sample of the individual's behavior appearing at one sitting was consistent with that at another time.

It was likewise thought that ten pictures might be too small a number of components for a test instrument to elicit a good sampling of behavior. Behavior under the test conditions might not be sufficiently extensive to give a satisfactory and reliable score.

Finally, it was observed from preliminary work with the test that the scoring was highly subjective and might be too dependent upon the "intuition" of the examiner at the expense of its reliability.

All these factors had to be taken into consideration, in studying the reliability of the ink-blot test.

3. Subsequent investigations of the Rorschach method are summarized by Beck (2, 3), Vernon (22), and Hertz (8).

Reliability is generally reported in terms of the coefficient of correlation, i.e., the correlation of the scores of the same individuals upon two successive and similar tests. In the absence of duplicate tests either the same material is used twice or the split test method is employed in which the test is divided into halves as comparable as possible, the score determined on each half, and those half scores correlated thus giving the reliability of half of the test.

### *Methods Used in Other Investigations*

Two studies are reported in which the Rorschach blots were given twice. Mira (11) applied the test to a group of subjects twice a fortnight apart. In certain groups he observed consistency of response. He considered the amount of change indicative of the stability of the individual. He does not include any statistical evidence of these findings. In like manner, Wertham and Bleuler (25) gave two applications of the test to determine differences in reactions of individuals under normal circumstances and when under the drug, mescaline. They reported comparatively close agreement in the two sets of responses, with one exception. The slight differences which did occur, did not materially affect the interpretation as a whole. In this study, also, no statistical computations are presented.

Behn-Eschenburg (4) reported the use of a parallel series of ink-blots which gave the same results as the standard Rorschach series. His claim is not, however, statistically substantiated. Other series of blots supposed to be similar to the original Rorschach blots have been used by Roemer (16, 17), Struve (21), Gordon and Norman (7) and Weil (23). These have not, however, been standardized against the Rorschach series, and there is no statistical evidence that they are actually similar in nature, equally difficult, and basically measuring the same factors.

The only statistical study to the writer's knowledge which specifically deals with the reliability of the method is that of Vernon (22). With ninety subjects, twenty-five male students of Yale University, forty-eight male students of Harvard, and seventeen male and female adults in England, he used the corrected split-half method and tested the reliability of the Rorschach categories. He divided each subject's scores into two halves of five blots, I, III, V, VI, and X in one series, and II, IV, VII, VIII, and IX, in a second series, correlated them and used the Spearman-Brown correction to show the predicted correlation of the whole test with a second similar test. Scores for his three groups of subjects were tabulated separately. He obtained the highest correlation for the number of responses ( $R$ ) + 0.91, indicating that if a parallel series were available, the average subject might attain the same



total number of responses as on the original set. The reliabilities of the other scores were not as satisfactory, the per cent W (see Table II for symbols) being 0.74, the percentage of M, O, and P falling between 0.70 and 0.60, and the other scores less, indicating that scores might be very different for the two sets. (See Table IV.) Vernon ascribes this unsatisfactory reliability of the different categories especially to the subjectivity of the scoring and to the shortness of the test. Because of this latter failing he recommends a parallel set or additional blots. He further states that psychometric standards of reliability do not "properly apply to the test" because Rorschach never intended each separate category to be isolated and treated as a quantitative variable. On the contrary he specifically insisted that no diagnosis should be made without considering the total psychogram. Vernon admits, however, "that if the test is to be regarded as a test at all, or if it claims any objective validity, it must in the future be modified in such a way that the reliabilities of the chief categories of response may achieve a level of at least 0.70-0.80."<sup>4</sup>

In the investigation of the Rorschach Test planned by the Brush Foundation<sup>5</sup> the split-test method was used in the preliminary experiment described in a previous paper.<sup>6</sup> Briefly, the test was administered to seventy subjects constituting a miscellaneous "normal" group, to thirty patients referred to the Psychological Clinic of Western Reserve University, and to fifty patients from the Neuropsychiatric Clinic of Lakeside Hospital, Cleveland, in two sittings, a set of odd-numbered blots being given at one time and after an interval of one to two weeks, the set of even-numbered blots. The answers were scored for those general test factors which required no norms or frequency tabulations, namely for the per cent W, per cent D (any detail answers), percentage of DS, M, C (any color answers), and per cent F (any form answers). The test factors were studied without their qualifications and their differentiations. The corrected coefficients for these factors showed a fairly high reliability in the three groups. Table I reproduces the reliability coefficients for the factors studied.

It was because the test in this general form showed satisfactory reliability that the experiment proper was set up. It was thought that if the procedure could be made more uniform and the scoring more objective, the reliability of the test could be definitely established.

4. Vernon, P. E. The Rorschach Ink-blot Test. *British J. Med. Psychol.*, XIII, Part III, 1933, p. 184.

5. Research was begun by the writer while a Fellow in Psychology at the Brush Foundation, Western Reserve University, and continued under the direction of Dr. L. Dewey Anderson. The Foundation is now called the Developmental Health Inquiry of the Associated Foundations.

6. Hertz, M. The Standardization of the Procedure and of the Scoring of the Rorschach Ink-blot Test. Western Reserve Library, Cleveland, Ohio.

TABLE I

Reliability Coefficients for Rorschach Scores with Correction\*

	No. of S.	%W	%D	%DS	%M	%C	%F
Miscellaneous	70	.66	.72	.60	.58	.62	.86
correction*		.80	.83	.75	.73	.77	.91
Psychological Clinic	25	.71	.69	.60	.24	.36	.55
correction*		.83	.82	.75	.39	.53	.71
Neuropsychiatric Clinic	50	.71	.61	.65	.55	.88	.81
correction*		.85	.76	.78	.71	.93	.69

$$* \text{Brown-Spearman Prophecy Formula } r_{12} = \frac{2r}{1+r}$$

### *Standardization of the Ink-blot Test*

Reference must be made to the previous paper for details as to standardization of the method of administering the test and of the method of scoring.<sup>7</sup> Suffice it to say here that revision was made of certain steps in the procedure to insure standardized conditions. Record blanks, summary sheets and diagrams were rearranged and improved, a trial blot was introduced into the procedure, time of exposure of each blot was limited to two minutes, uniform directions were given, and the test situation was kept as uniform as possible. The test was given to 300 students of Patrick Henry Junior High School, Cleveland, 150 boys and 150 girls, ages ranging from 12, 6 to 16, 5. The test records obtained were subjected to statistical analysis and definite quantitative and qualitative criteria were determined for scoring the Rorschach test factors. Responses were tabulated for each ink blot, the tabulations constituting "Frequency Tables." These tables represented the standard of normality for the specific age group and were used in scoring W, D, Dr, Do, F, O, I and P. The scoring method was found to be reliable as judged by the extent of agreement between the scores of one judge and those of another on 100 records selected at random.

### *Procedure in Determining the Reliability of the Test*

It was decided to use the split-test method again in determining the reliability of the ink-blot test as had been done in the preliminary work. Two

7. *Ibid.*



applications of the original set were not advisable since factors of familiarity and practice would enter and prejudice results. No parallel series of blots was available. The split-test method was the only method left.

This procedure has its limitations. The assumption is made that halves of the test are approximately equivalent in difficulty, content, and similar respects. When the Rorschach test was divided into a set of five odd-numbered cards (Set A) and another set of five even-numbered cards (Set B), it was observed that the latter contained one colored blot more than the former. It is to be noted that Vernon's sets (22) were not strictly comparable for the same reason. Second, there are several test factors of diagnostic significance which very obviously could not be studied with this method because their existence depends upon the original progression of the series, the *succession* for example. These factors had to be omitted from this study because of the method of approach.

Using the split-test method was justified because it seemed to be the only method available. No test factors which depended upon the progression of the series were included in this treatment. This method further would reveal the influence of the variability-performance factor, the adequacy of the instrument, and of the standardization of the procedure in reference to the test factors selected for study.

### *Computation of the Coefficients of Reliability*

One hundred records were selected at random from the group of three hundred. Scores of each of these were divided into two parts, Set A containing scores of the answers to the odd numbered cards and Set B containing scores of the responses to the even numbered cards. The two sets of scores were summarized. In all there were two hundred half sets of scores.

The two sets of scores were compared for the following factors:

1. Total number of responses
2. Total number of whole responses
3. Percentage of whole answers
4. Number of normal detail answers
5. Percentage of normal detail answers
6. Number of rare detail answers
7. Percentage of rare detail answers
8. Number of oligophrenic details
9. Percentage of oligophrenic details

10. Number of normal details, rescored and not considering the oligophrenic category
11. Percentage of normal details, rescored and not considering the oligophrenic category
12. Number of rare details rescored and omitting consideration of the oligophrenic category
13. Number of white space details
14. Percentage of white space details
15. Number of good forms
16. Percentage of good forms on the basis of the total responses
17. The %F+ factor of Rorschach
18. Number of movement answers
19. Percentage of movement answers
20. Number of chiaroscuro responses
21. Percentage of chiaroscuro responses
22. Number of color answers of any kind
23. Percentage of color answers of any kind
24. Color score factor of Rorschach
25. Number of animal answers
26. Percentage of animal answers
27. Number of human answers
28. Percentage of human answers
29. Number of answers referring to anatomy
30. Percentage of anatomical answers
31. Number of good original answers
32. Percentage of good original answers
33. Number of popular responses
34. Percentage of popular responses
35. Number of forms verbally mentioned, called "items"
36. The "Erlebnistypus" as suggested by each set

In Table II are presented the reliability coefficients of certain of these factors with correction according to the Brown-Spearman formula.

The Rorschach factors appear to be reliable in most cases. Percentage of anatomical, original, and chiaroscuro answers showed the highest reliability (.9). Satisfactory coefficients (.8) were obtained likewise for number of responses, percentage of whole, rare detail, oligophrenic detail, space detail, color, animal and human form answers, and number of items. Percentage of normal detail, good form, and movement answers and the color score obtained coefficients approximating .70. The lowest coefficient (.6) in the group was that for percentage of popular answers. The normal detail and



the rare detail factors appeared to have greater reliability when the oligophrenic detail was omitted. The oligophrenic detail, however, showed high reliability in itself.

Comparing *Erlebnistypen* suggested by each half of the test, the percentage of correspondence is 73.

### *Comparison with Results in Preliminary Experiment*

In the preliminary work, reliability coefficients were found for the percentage of whole, detail, space detail, movement and color answers, treated generally without consideration of qualifications. The split-test method was used, each half of the test being given at a separate sitting.

TABLE II  
Reliability Coefficients of Rorschach Test Factors ( $N = 100$ )

Sym.	Test Factors	Coeff.	P.E. $\pm$	Correction*
R	Total number of responses	+ .812	.014	+ .890
%W	Percentage of <i>whole</i> answers	+ .717	.032	+ .835
%D	Percentage of <i>normal detail</i> answers	+ .600	.043	+ .750
%Dr	Percentage of <i>rare detail</i> answers	+ .752	.029	+ .859
%Do	Percentage of <i>oligophrenic detail</i> answers	+ .686	.040	+ .813
%D	Percentage of <i>normal detail</i> answers (No Do)	+ .678	.039	+ .808
%Dr	Percentage of <i>rare detail</i> answers (No Do)	+ .814	.024	+ .897
%DS	Percentage of <i>space detail</i> answers	+ .767	.030	+ .868
%F	Percentage of <i>good forms</i> in relation to R	+ .677	.039	+ .807
%F+	Percentage of <i>good forms</i> in relation to F	+ .576	.051	+ .730
%M	Percentage of <i>movement</i> answers	+ .594	.043	+ .745
%F(C)	Percentage of <i>chiaroscuro</i> answers	+ .847	.024	+ .917
%C	Percentage of <i>color</i> answers	+ .677	.039	+ .810
C	<i>Color Score</i>	+ .640	.039	+ .763
%A	Percentage of <i>animal</i> answers	+ .708	.035	+ .829
%H	Percentage of <i>human</i> answers	+ .756	.029	+ .861
%Anat.	Percentage of <i>anatomical</i> forms	+ .944	.013	+ .971
%O+	Percentage of <i>original</i> answers	+ .837	.024	+ .911
%P	Percentage of <i>popular</i> answers	+ .495	.051	+ .666
Items	Total number of <i>Items</i> given	+ .761	.017	+ .864

Percentage of correspondence between  
*personality types* 73%

\* Brown-Spearman Prophecy Formula  $r_{12} = \frac{2r}{1+r}$

The reliability coefficients with the final results here were for the same factors in more refined form. The same procedure was used except that the test was given in one sitting.

The two sets of coefficients are therefore not comparable. However, they are placed side by side in Table III for observation.

The reliability of the whole, space detail and color answers appear to have been increased. The detail factor is not at all comparable of course, since in the final experiment it referred to normal details as statistically determined whereas in the preliminary work it meant any detail answer.

TABLE III

Coefficients of Reliability (as Corrected) in Preliminary and Final Experiments

Group	No. of S.	%W	(%D)	%DS	%M	%C
<i>Preliminary</i>						
Normal	70	.80	(.83)*	.75	.73	.77
<i>Final</i>						
Patrick Henry	300	.89	(.80)*	.87	.76	.81

\* Not comparable.

### *Comparison with Results Obtained by Vernon (22)*

Table IV reproduces the correlations found by Vernon, averaged for his three groups. It is to be noted that his reliabilities are considerably lower than those obtained in this investigation. His average is 0.54 while the coefficients in the present study average 0.829. The number of responses and the percentage of whole answers appear to be reliable in both studies. The percentage of original answers given irrespective of qualification as used in Vernon's study has a much lower reliability (0.60) than the percentage of good original answers as herein obtained (0.911). The percentage of popular answers obtains unsatisfactory reliability in both studies. It is to be noted that the %F+ in the present study has a much higher reliability (.73) than the same factor in Vernon's study (.33). This is likewise true of the %A which has a reliability of .829 here while it attained only .48 in the other study. Again the color score appears more reliable (.763) than the sum color score per cent of Vernon.

Possibly the standardization of the procedure, especially limitation of the time of exposure, and standardization and increased objectivity of the scoring can account in part for the better results herein obtained.



TABLE IV

Vernon's Reliabilities of Scores on the Rorschach Test ( $N=90$ )

Test Factors	Average for Three Groups
W%	0.74
F+%	0.33
M%	0.62
SumC%*	0.34
A%	0.48
O%*	0.60
P%	0.64
Aver.	0.54
R	0.91

\* Note—It is to be noted that these factors are not comparable to similar factors in the present investigation. Color score here is on a percentage basis, whereas in this study, only the percentage of color answers given and the sum color score were used. Again, Vernon used the percentage of original answers given while the percentage of good original answers is herein employed.

In determining the reliability of the *Erlebnistypus*, Vernon used the formula  $\frac{M - \text{Sum C}}{R}$  as an index of the degree of introversion and extra-intension. He obtained an average split-half reliability (corrected) of 0.55, a result which did not bear out Rorschach's contention that the single test is adequate to diagnose the personality type. According to our results, a correspondence of 73 per cent was obtained between the types found on each half of the test, which tended to confirm Rorschach.

### *Summary of the Reliability of the Rorschach Test*

1. The coefficients of reliability of the test factors obtained by the corrected split-test method may be considered satisfactory, since they compare favorably with most of the correlations reported in the literature for intelligence tests or tests of mechanical abilities.

2. The ten component pictures of the Rorschach series are sufficient to elicit a fair sampling of an individual's behavior or personality in terms of the above test factors.

3. Personality traits or behavior patterns in terms of the above factors tend to be consistent and follow a stable and unitary pattern. Hence these patterns may be considered as an accurate index to the individual's habitual responses to the test, and the variability performance factor, though it ap-

pears important, evidently does not militate against the reliability of the test.

4. Comparison of the indices of reliability in the preliminary experiment where some general test factors were considered, and in the final work where test factors were studied in their several differentiations and where there was refinement of technique, indicates that the reliability of the test was increased after the procedure was standardized and the scoring was made more objective.

### *Intercorrelations of the Rorschach Test Factors*

Despite the fact that Rorschach does not introduce any statistical data in his manual, he does indicate certain definite relationships between the various test factors. An examination of these would serve as a further check on the reliability of the method.

Behn-Eschenburg (4), already referred to, compared the average scores of each test factor of his extreme groups to ascertain if Rorschach's relationships obtained. He did not use the method of correlation. He reported positive correspondence between W and M, per cent F+ and FC, also between per cent F+ and Mr; negative correspondence between W and Dr, per cent A and M, Dr and DS, and between per cent F+ and Dr.

Vernon (22) studied the interrelations of the different test factors by the method of correlation and reported statistically unreliable results. He again attempted to account for these unsatisfactory results in terms of Rorschach's assertions that the relationships are to be found only in the "normal person" and are at times broken down or even completely reversed by emotional disturbances. Since the normal person according to Rorschach is the hypothetical person who never can be found (if he was, he would be abnormal) one should not expect to find the schematic relationships enumerated by him.

Table V contains the correlations which obtain among different factors selected for study in the present investigation.<sup>8</sup>

Comparing the whole answer factor with the others, the highest relationship appears between the movement and the color score and the percentage of good original answers. It corresponds least with the percentage of good forms. Rorschach found the direct proportion between the whole an-

8. For treatment of the relationship between the Rorschach test factors and intelligence, see Hertz, M. *Validity of the Rorschach Test: Intellectual Factors*. Western Reserve University, Library, Cleveland, Ohio.



TABLE V  
Intercorrelations\* of the Rorschach Test Factors (N = 300)

	IQ	%F+	%O+	Items	%A	Color Score	Whole Ans.	Move. Ans.
IQ		.460	.398	.186	-.108	.209	.241	.259
%F+	.460		.263	-.003	-.007	.072	.108	.173
%O+	.398	.263		.390	-.436	.305	.380	.509
Items	.186	-.003	.390		-.281	.294	.240	.482
%A	-.108	-.007	-.436	-.281		-.311	-.246	-.294
Color Score	.209	.072	.305	.294	-.311		.393	.177
Whole Ans.	.241	.108	.390	.240	-.246	.393		.424
Move. Ans.	.259	.173	.509	.482	-.294	.177	.424	

\* P.E. of the coefficients of correlation for r

0.0 - 0.2 ± .0380

0.3 - 0.5 ± .0323

swer and the movement which is suggested here.<sup>9</sup> He did not find clear proportions between W and F or between W and C.<sup>10</sup> Present data agree in that they do not indicate high correspondence between these factors. Experience with the test bears out the fact that W need not necessarily be associated with good form imagery. Subjects frequently gave whole answers without consideration of the exactness of the forms involved. This is pointed out in the study of the feeble-minded group<sup>11</sup> who, of all the groups studied, gave the highest percentage of whole answers with the lowest percentage of good forms. Again, many normal subjects gave few whole answers with a high percentage of good forms. Data here presented suggest that the more whole answers given, the less the stereotypy and the more originality and productivity, which again is in accord with Rorschach's finding.

The movement factor appears to vary directly with the percentage of good original answers and with the number of items. This appears to be in accord with Rorschach's estimate of this factor.

"The number of M rises with the productivity of the intelligence, with the wealth of associations, and with the ability to complete new associative connections."<sup>12</sup>

"A direct proportion exists between movement and original answers."<sup>13</sup> Further, a low negative correlation is likewise obtained with the percentage

9. Rorschach, H. *Psychodiagnostik*. Pp. 30 and 54.

10. *Ibid.*, p. 30.

11. Hertz, M. *Validity of the Rorschach Test: Intellectual Factors*. Western Reserve University Library, Cleveland, Ohio.

12. Hertz, M. *Validity of the Rorschach Test: Intellectual Factors*. Western Reserve University Library. P. 15.

13. *Ibid.*, p. 80.

of stereotypy, again showing the adverse relationship suggested by Rorschach.

"Stereotyped and feeble-minded persons have no movement answers."<sup>14</sup>

"The number of movement answers stands in direct proportion . . . to the variability of the responses, hence in reverse proportion to the animal per cent. And the direct proportion between the number of good original answers and the movement answers stand out even more clearly."<sup>15</sup>

The coefficient of correlation obtained between M and per cent F+ (+.173) does not substantiate Rorschach when he reports:

"With normal subjects, a clear proportion exists between the number of M and the keenness of the form imagery."<sup>16</sup>

Rorschach adds, however, that this direct proportion may be upset whenever emotional disturbances enter.<sup>17</sup> In like manner, Rorschach is not confirmed when he indicates a direct proportion between movement and color factors, but again he indicates exceptions.<sup>18</sup>

Rorschach indicated a proportion between per cent F+ and C,<sup>19</sup> but practically no relationship is indicated by present data. Again, Rorschach finds this correspondence may be reversed by normal subjects with nervous or artistic leanings.<sup>20</sup> The inverse relationship between color and stereotypy indicated by Rorschach is confirmed by the correlation of  $-.311$  which was obtained.

"Only the stabilized show no color . . . the less color, the more stabilized and the more stereotyped the individual."<sup>21</sup>

Present data also indicate a positive relationship of  $+.305$  with the per cent 0+ which bears this out.

The percentage of stereotypy appears negatively related to per cent 0+, to the number of items, the color score, whole and movement answers. The highest relationships obtained are with the per cent 0+ ( $-.436$ ) and the color score ( $-.311$ ). These facts are all suggested by Rorschach. He indicates that artistic and imaginative people have a low per cent A, while the stereotyped and the feeble-minded have a high per cent A.<sup>22</sup>

It should be noted that the number of items shows highest correlation with the per cent 0+ and the movement answers. This factor was introduced into the present study and was not suggested by Rorschach.

14. *Ibid.*, p. 15.

15. *Ibid.*, p. 55.

16. *Ibid.*, p. 17.

17. *Ibid.*, pp. 17 and 54.

18. *Ibid.*, p. 21.

19. *Ibid.*, p. 21.

20. *Ibid.*, p. 21.

21. *Ibid.*, p. 22.

22. *Ibid.*, p. 36, Table VI.



### *Summary for the Intercorrelations of the Rorschach Test Factors*

1. Intercorrelations range from  $-.436$  to  $+.509$ .
2. The whole answer factor shows highest correspondence with movement, color score, and percentage of good original answers.
3. The movement factor is best related to the good original answers and corresponds likewise to items and number of whole answers.
4. The color score is positively related to per cent 0+ and W and negatively related to per cent A.
5. The per cent A shows a negative relationship to all the other factors especially with the per cent 0+, the C and the W.
6. The items factor is best related to movement and originality.
7. While definite conclusions cannot be made, the figures suggest many of the findings which Rorschach reported.

### *General Conclusion*

Statistical treatment of the results obtained with the Rorschach test in its modified and standardized form shows the test to be a reliable instrument.

### *BIBLIOGRAPHY*

1. BECK, S. J. Personality Diagnosis by Means of the Rorschach Test. *Amer. J. Orthopsychiat.*, 1930, 1, 81-88.
2. ———. The Rorschach Test and Personality Diagnosis. I. Feeble-minded. *Amer. J. Psychiat.*, 1930, 10, 19-52. Also in *Institute for Child Guidance Studies* (L. G. Lowrey), pp. 222-261. N. Y.: Commonwealth Fund, 1931.
3. ———. The Rorschach Test as Applied to a Feeble-Minded Group. *Archives of Psychol.*, 1932, 136, 84.
4. BEHN-ESCHENBURG, H. *Psychische Schüleruntersuchungen mit dem Formdeutversuch*. Ernst Bircher Verlag, Bern u. Leipzig, 1931, p. 69. Also Huber, Bern.
5. ENKE, W. Die Konstitutionstypen im Rorschachschen Experiment. *Zsch. f. d. Neur. u. Psychiat.*, 1927, 108, 645-674.
6. ———. Die Bedeutung des Rorschachschen Formdeutversuches für die Psychotherapie, *Sitzungsber. d. Ges. z. Beförderung, d. ges. Naturw. z. Marburg*. 1927, 62, 621-633, Berlin: O. Elsner Verlag.
7. GORDON, R. G. AND NORMAN, R. M. Some Psychological Experiments on Mental Defectives in Relation to the Perceptual Configurations Which May Underlie Speech. Part II. *Brit. J. Psychol.*, 1932, 23, 85-113.

8. HERTZ, M. R. The Rorschach Ink-blot Test. Unpublished thesis, Western Reserve University Library, Cleveland, Ohio.
9. LOOSLI-USTERI, M. Le test de Rorschach, appliqué à différents groupes d'enfants de 10-13 ans. *Archives de Psychol.*, 1929, 85, 51-106.
10. LOPFE, A. Über Rorschachsche Formdeutversuche mit 10-13 jährigen Knaben. *Zsch. f. angew. Psychol.*, 1925, 26, 202-253.
11. MIRA, L. E. Sobre el valor del Psicodiagnostico de Rorschach. *Progressos de la Clinica*, 1925, 808-845.
12. MÜLLER, M. Der Rorschachsche Formdeutversuch, seine Schwierigkeiten und Ergebnisse. *Zsch. f. d. ges. Neur und Psychiat.*, 1929, 118, 598-620.
13. MUNZ, E. Die Reaktion des Pykniques im Rorschachschen psychodiagnostischen Versuch. *Zsch. f. d. ges. Neur. und Psychiat.*, 1924, 91, 26-92.
14. OESER, O. A. Some Experiments on the Abstraction of Form and Color. Pt. II. Rorschach Tests. *Brit. J. Psychol.*, 1932, 22, 287-323.
15. PFAHLER, G., System der Typenlehren. Grundlegung einer pädagogischen Typenlehre, *Zsch. f. Psychol.*, 1929, Erg. Bd. 15, p. 334.
16. ROEMER, G. A. Die Innenwelt einer Persönlichkeit und das Problem ihrer wissenschaftlichen Erschliessung. *Psychol. Rundschau*, 1930, 2, 4-12. *Psychol. Rundschau*, 1930, 2, 33-41. *Ibid.*, 69-76; 101-109. Reprint: *Die wissenschaftliche Erschliessung der Innenwelt einer Persönlichkeit*, Basle: Birkhäuser, 1931, p. 42.
17. ———. Psychographische Tiefenanalyse einer Grossindustriellen und seiner Stabes. *Prakt. Psychol.*, 1922-3, IV, 16-28.
18. RORSCHACH, H. *Psychodiagnostik. Methodik und Ergebnisse eines wahrnehmungsdiagnostische Experiments*. Bern: Bircher, 1921, p. 42. Also enlarged, Bern: Huber, 1932, p. 227.
19. RORSCHACH, H. AND OBERHOLZER, E. Zur Auswertung des Formdeutversuche für die Psychoanalyse, *Zsch. f. d. ges. Neur. und Psychiat.*, 1923, 82, 240-273. The same, translated: The Application of the Interpretation of Form to Psychoanalysis. *J. Nerv. and Ment. Dis.* 1924, 60, 225-248; 359-379.
20. SOUKUP, F. The Study of Personality and Rorschach's Test. In Czech. *Čas. lék. česk.*, 1931, 1, 881-887. Abstract: *Zentbl. f. d. ges. Neur.*, 1932, 61, 564.
21. STRUVE, K., Typische Ablaufsformen des Deutens bie 14- bis 15- jährigen Schulkindern, *Zsch. f. angew. Psychol.*, 1930, 37, 204-274.
22. VERNON, P. E. The Rorschach Ink-blot Test. *Brit. J. Med. Psychol.*, 13, 1933, 89-114; 179-200, 271-291.
23. WEIL, H. Wahrnehmungsversuche an Integrierten und Nichtintegrierten. *Zsch. f. Psychol.*, 1929, 111, 1-50.
24. WELLS, F. L. The Systematic Description of Personality. Conference on Individual Differences in the Character and Rate of Psychological Development. Washington: National Research Council 1931, pp. 52-70.
25. WERTHAM, F. AND BLEULER, M. Inconstancy of the Formal Structure of the Personality: experimental study of the influence of mescaline on the Rorschach test. *Arch. Neur. und Psychiat.*, 1932, 28, 52-70.

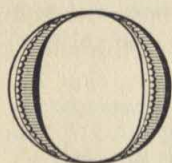


Richard H. Dana

## RORSCHACH SCORER

### RELIABILITY<sup>1</sup>

#### *Problem*



OVER A thirty-year period the Rorschach technique has enjoyed an ever increasing frequency of application within a continually expanding clinical and research framework. Among the most important and least investigated aspects of Rorschach testing is the problem of reliability. There are at least two kinds of reliability germane to projective instruments including (a) reliability of the instrument and (b) reliability of the scorer. The literature contains discussions and research on reliability of the test *per se* but there are few published reports of scorer reliability (1, 6, 7). In a standard reference, this problem is only briefly mentioned: "It is desirable to make the system of classification as reliable as possible, so that different examiners will arrive at essentially similar classifications of the same sets of responses" (5, p. 19). In a recent overview, Hertz (4) does not mention reliability. Considering the importance of individual scores in the psychogram and the ratios calculated,

Reprinted from *J. Clin. Psychol.*, 1955, 11, 401-403, by permission of the publisher and the author.

1. This investigation was supported by a research grant from the National Institute of Mental Health, of the National Institutes of Health, Public Health Service. Principal Investigators are: A. G. Ossorio and A. K. Busch. Research Associate: R. H. Dana. Acknowledgment is made to Jerome Pauker and Jack Werboff, Research Assistants, who did the reliability scoring.

the problem of scorer reliability should have received more attention in the past.

The paucity of research in this area may be due to an un verbalized awareness that conventional reliability measures may not be applicable to projective materials. A beginning has been made in discussions of reliability in TAT scoring (2), content analysis (8), and more generally, in projective techniques (3). It is sufficient to indicate here that per cent of agreement, which involves no assumptions, may have more applicability than the usual product moment or tetrachoric correlations between scores of two independent scorers.

### *Method*

A representative sample of Rorschach records for each of eleven examiners was obtained by (a) following an alphabetized list of examiners, one record was pulled for each examiner from consecutively numbered files (numbered by admission date) until eleven records had been drawn; (b) this process was repeated six times for a total sample of sixty-six records; (c) this method permitted sampling from all the records in the files for a four-year period.

These records were scored independently by two scorers of approximately equal training and experience in other locales, using the Klopfer method (5). Per cent of agreement was applied as follows: (a) location, determinants, content, and popular main scores were treated as separate items. For example, the response W M H P would consist of four separate items; (b) the total N of separate items present in the six scored records for each examiner was computed; (c) using the N of examiner items as criterion, comparisons were made for presence or absence of identical items in the scoring of the independent scorers.

### *Results*

Thirty-three per cents of agreement were computed (Table 1): Scorer I, Scorer II, and Scorer I with Scorer II, for each of the eleven examiners.

### *Discussion*

There was a mean difference of 4.4 points between the agreements of Scorers I and II. This difference is attributable to *degree of care exercised*



TABLE 1

Rorschach Per Cents of Agreement Between Two Scorers and Eleven Examiners

Examiners	N Possible Agreements	Scorers		
		I	II	I, II
1	235	80	86	86
2	455	73	76	73
3	324	77	84	82
4	414	69	74	75
5	328	71	77	77
6	394	74	74	82
7	291	64	74	75
8	425	71	75	69
9	299	77	78	82
10	409	73	74	75
11	307	69	75	73

and use of *personalized scoring variables*. Inspection of all scores indicated that these two kinds of errors, on the part of both the original examiners and the reliability scorers, account for most of the disagreement present.

A figure of 75 per cent of agreement can be considered representative of scoring reliability in this hospital situation. The sample of Rorschach records was scored with adequate reliability to justify inclusion of the entire population of Rorschachs by these same examiners in any research requiring use of Rorschach scoring categories. It is suggested that this is an economical method of determining whether or not there exists a need for re-scoring when hospital Rorschach records are used for research purposes.

This method may not have been applied previously because of the implicit assumption that there is only one *right* way of scoring each response and that equally well trained scorers will score in exactly the same manner. The eleven original examiners in this study had from one to seven years of Rorschach experience; the reliability scorers each had four years of experience. There were no significant relationships between years of experience and reliability of scoring.

The only other use of per cent of agreement for Rorschach reliability reported an over-all figure of 76 per cent using an N of 300 responses, five internationally renowned scorers, and a criterion of agreement by three scorers (7). This striking agreement between experts and the present hospital examiners suggests that increase in training and experience beyond a certain degree of competence does not appreciably affect scorer reliability. If future studies replicate a figure of approximately 75 per cent of agreement for Rorschach scoring, then the 25 per cent of disagreement may be indicative

of ambiguity of the scoring categories or of personalized scoring, i.e., influences of examiner personality.

### Summary and Conclusions

1. Scorer reliability for two independent scorers with each of eleven Rorschach examiners was computed by means of per cent of agreement.
2. The thirty-three agreement figures ranged from 64 per cent to 86 per cent with 75 per cent representative of this clinical situation.
3. These per cents of agreement appear sufficiently high for research purposes and the extent to which they fall short of optimal agreement may be a function of ambiguous scoring categories or personalized scoring.
4. Scorer reliability is considered an economical method of determining the need for rescoring Rorschach records to be used in research.
5. The usefulness of per cent of agreement as an index of scorer reliability appears to have been overlooked in objectification of the Rorschach and other projective instruments.

### BIBLIOGRAPHY

1. AMES, L. B., LEARNED, J., METRAUX, R. W. AND WALKER, R. N. *Child Rorschach Responses*. New York: Paul B. Hoeber, Inc., 1952.
2. DANA, R. H. Clinical diagnosis and objective TAT scoring. *J. Abnorm. Soc. Psychol.*, 1955, 50, 19-25.
3. DANA, R. H. The objectification of projective techniques: Rationale. *Psychol. Rep.*, 1955, 1, 93-102.
4. HERTZ, MARGUERITE. The Rorschach: thirty years after. In Brower, D. and Abt, L. (Eds.) *Progress in Clinical Psychology*. Vol. I. New York: Grune & Stratton, 1952.
5. KLOPFER, B., AINSWORTH, MARY, KLOPFER, W. AND HOLT, R. *Developments in the Rorschach Technique*. Vol. I. Yonkers-on-Hudson: World Book Co., 1954.
6. RAMZY, I. AND PICKARD, P. M. A study in the reliability of scoring the Rorschach ink-blot test. *J. Genet. Psychol.*, 1949, 40, 3-10.
7. SICHA, K. AND SICHA, M. A step towards the standardization of the scoring of the Rorschach test. *Ror. Res. Exch.*, 1936-37, 1, 95-101.
8. SPIEGELMAN M. TERWILLIGER, C. AND FEARING, F. The reliability of agreement in content analysis. *J. Soc. Psychol.*, 1953, 37, 175-187.



*Bert Kaplan*  
*Stanley Berger*

*INCREMENTS AND CONSIST-  
ENCY OF PERFORMANCE IN  
FOUR REPEATED RORSCHACH  
ADMINISTRATIONS*

**I**N WORK with the Rorschach, interpretations are almost always based on a single test administration. The assumption is made that only in his initial and spontaneous reactions does the subject select responses which optimally reveal his personality and that these responses, with a few additions, constitute the entire universe of psychologically meaningful or interesting responses. The basis for this assumption undoubtedly lies in the theory that perception of the Rorschach blots is selective and that what is initially selected is of special significance. In the present paper, we will take the position that this selective process continues well beyond the perceptions which are ordinarily reported on a first administration of the

Reprinted from *J. Proj. Tech.*, 1956, 3, 304-309, by permission of the publisher and the authors.

Rorschach and that the development of a technique for eliciting large numbers of additional perceptions is of psychological interest.

A number of considerations lead us to question whether it is wise to put all our eggs in the basket of a single administration of the Rorschach test. Experience in psychotherapy and in personality study indicates that very often patients are unwilling or unable to express their major concerns and preoccupations on an initial contact and only very gradually, after many sessions, do they become freely expressive. We suggest that many of the instances in which the Rorschach "does not test" as Lowell Kelly has put it, or does so in a disappointingly sparse way, are cases of this kind of initial reticence or inhibition rather than "emotional impoverishment," "shallowness," or "insufficient differentiation."

In this connection, a host of recent studies, such as those by Lord (7), Kimble (5), Klatskin (6), Eichler (1), Gibby (2) and Hutt and his colleagues (3) indicate that the circumstances surrounding the administration of the test have a considerable influence on the nature of the responses. This influence is not superficial but affects the most basic aspects of the performance. If test results derive in part from the specific situation in which the subject finds himself at a particular moment, we may draw the conclusion that any one test performance can hardly tell the whole story, since in other situations the subject might appear to be quite different.

The most convincing argument against the complete reliance of the Rorschach worker on a single performance lies in the astonishing complexity of personality. As clinicians, sensitive to this fact, we expect that even when an interview or test procedure yields apparently profound insights into personality, subsequent information will broaden the picture and modify our impressions. This process seems to go on as long as we continue to study the subject and we literally never achieve a final or true formulation. Work at the Harvard Psychological Clinic with the TAT has shown that only part of the potential of the subject is tapped by the twenty pictures in the test and that with repeated administrations of the TAT or with additional testing of the storytelling variety, important additional material can be obtained which not only expands our picture of the subject, but often sharply alters any formulation based on the first administration.

Retest studies of Rorschach performance usually reveal a great deal of similarity in responses from one administration to another. Even when there have been drastic intervening conditions, the stability of perceptions is remarkable. We believe that this stability is a function of the dominance of what has already been seen and that perceptual organizations once achieved tend, under certain conditions, to interfere with the emergence of new ones. The conditions we have in mind exist when the old perceptions



are in competition with potential new ones. This is the case in the usual retest situation. If we ask why the old perceptual organizations do not interfere with new ones during the course of a single test we can say that a response which has already been given does not compete with new ones which are emerging since only the new responses satisfy the demands of the test situation. The older responses are no longer of value in terms of the immediate needs of the subject and thus lose their compelling and dominating force.

Since we were, in the present study, interested in seeing whether substantial numbers of additional responses could be produced, this analysis offered a clue to the appropriate method for eliciting them. This was simply, on the second and subsequent administrations, to ask the subject to report only new perceptions and to omit anything he had given earlier. This instruction we hoped would remove the dominating and inhibiting influence of what had already been seen.

Using the instruction to report only new and different responses, the Rorschach was administered four times. We had the following questions in mind: 1) Can subjects produce a substantial number of responses in addition to the ones given on a first performance? 2) Do these new responses, if forthcoming, add significantly to the personality picture derived from a first performance? 3) Are the new responses, if forthcoming, substantially similar to the ones given on a first performance in terms both of content and scoring categories? We believed that the answers to these questions would have certain implications for Rorschach practice and theory, for the question of the reliability of the test and would yield new insights into the correlates of Rorschach productivity, a subject dear to the heart of anyone who has given three fifteen-response Rorschachs in succession.

### *Procedures*

The subjects in this study were twenty-eight college students who were volunteers from an introductory psychology course at the University of Kansas. The group Rorschach was administered four times, using the Harrower-Erickson slides which were projected on a screen and exposed for a minute and a half in the regular position and a minute and a half in an inverted position. On the second, third and fourth administrations, the subjects were told to give only new and different responses. The protocols were subsequently examined and a very small number of responses which we judged to be duplicates were eliminated. The time interval between the tests was four days. A special effort was made to create and maintain a

friendly and relaxed atmosphere and high group morale throughout the experiment. A fifth session was held in which an inquiry was conducted, and a sixth session was held in which the subjects filled out a questionnaire asking for certain attitudes and feelings with respect to the test experience. The experiment was also explained and discussed during this session. It is of interest that all twenty-eight of the subjects who started continued in the experiment until the last session, which was optional.

## Results

Table I presents the mean scores for seventeen Rorschach variables on each of the four administrations. It may be seen that productivity as measured by R fell off about a third on the second administration, but maintained itself at the same level on the third and fourth. To our question as to whether subjects can produce substantial numbers of new responses, we can answer in the affirmative since subjects who gave an average of thirty-four responses originally, produced an average of 68.6 new responses on the three later performances. These results also strongly suggest that this is not the outer limit of productivity but that a fifth and sixth administration would bring even more responses. Informal experiments with as many as nine administrations have revealed there is a gradual decrease both in number and quality.

TABLE I  
Mean Score of 28 Subjects in 4 Repeated Rorschach Administrations

	Test I	Test II	Test III	Test IV
R	34.8	23.9	21.0	23.7
M	7.2	5.4	4.5	4.1
FM + m	9.1	5.9	5.1	5.1
k + K	.8	.3	.3	.4
FK	.3	.3	.2	.2
F	11.9	8.9	7.9	8.6
Fc + c	3.4	1.7	1.5	1.8
C'	1.6	1.5	.9	1.4
FC	5	3	2.9	2.5
CF + C	1.2	.9	.5	1.4
sum C	3.9	2.5	2.3	4.6
M:sum C	1.8	2.2	1.9	.9
W	17.8	9.4	7.4	12
D	14.7	12.9	9.6	8.7
d	.2	.4	.4	.4
Dd + s	2.5	2.6	3.3	2.6



Table II presents Kendall's coefficient of concordance, a non-parametric test of consistency of rank, for thirteen Rorschach variables. It may be seen that twelve are significant at the .01 level and one at the .02 level. These findings indicate that there is a significant degree of stability in the scores on the location and determinant categories over the four performances. We should mention that these scores were in the form of percentages rather than numbers of responses, so that we were working with M %, FC%, etc. This means that the fact that individuals tended to hold their ranks in all four

TABLE II

Kendall's Coefficient of Concordance, W, Showing Stability of Ranks on 13 Rorschach Variables of 28 Subjects in 4 Administrations of the Rorschach Test

Variable	"W"	p
F	.49	.01
M	.46	.01
FM + m	.48	.01
Fc + c	.50	.01
FC	.42	.01
sum C	.40	.01
M:sum C	.37	.02
W	.65	.01
D	.50	.01
Dd and S	.52	.01
Sum c*	.48	.01
M:sum C*	.42	.01
R*	.64	.01

\* On the starred measures, "W" is based upon actual frequencies of responses. For those which are not starred, the scores were converted to percentages.

performances was not a function of differences in the number of responses. These findings do not, however, resolve adequately the question of whether the responses in each of the four performances are distributed in the same way among the determinant and location categories. The significant W's indicate that a similarity does indeed exist but a question remains as to the magnitude of the similarity and it is difficult to translate the size of the W's into such terms.

The product-moment correlation provides an index of similarity of relationship between sets of scores. Table III presents the correlations between Tests 1 and 2, 1 and 3, and 1 and 4 of scores on 13 variables. Correlations are given both for scores converted to percentages as in M % and for scores reflecting simple frequencies as in number of M responses. Of the forty-three correlations computed, seventeen are significant at the .01 level and 3 others are significant at the .05 level. R, FM, and M, and the location cate-

gories tend to show the greatest amount of consistency. The R's, taken as a group, indicate that there is a considerable amount of relationship between performance I and subsequent performances. However, the coefficients of alienation indicate that only a relatively small part of the variance is accounted for by these correlations. The highest correlation, .75 for W% between Tests 1 and 2 accounts for 52% of the variance while an R of .51 accounts for only 27% of the variance. The findings indicate therefore that there has also been a considerable amount of change. These changes were large enough so that the basic shape of the psychogram showed relatively little stability. Using only percentages of movement, form and Sum C responses it was found that when the patterns formed by these scores were compared for Tests 1 and 2, 1 and 3, and 1 and 4 for the 28 subjects separately, on only 18% of the pairs was the pattern the same so that for example, F% was still highest, M% second and Sum C% lowest.

Perhaps the most interesting data in this study have to do with the changing content of responses in the four administrations. Unfortunately space does not permit our going into this in more than a cursory way. In

TABLE III  
Product-Moment Correlations of Rorschach Scores of 28 Subjects in 4  
Repeated Administrations

Test	1 and 2		1 and 3		1 and 4	
	Freq.	%	Freq.	%	Freq.	%
R	.72**		.56**		.64**	
M	.07	.51**	.41*	.17	.33	.06
FM + m	.49**	.41*	.60**	.57**	.54**	.53**
F		.54**		.36		.17
FC	.30	.28	.38	.41*	.13	.37
Sum C	.16	.32	.20	.21	-.04	.20
M:Sum C	.16		.53**		.22	
W		.75**		.70**		.49**
D		.67**		.46		.62**
Dd and S		.54**		.47		.08

\* Significant at the .05 level.

\*\* Significant at the .01 level.

order that the reader get some idea of the concrete situation, however, we have presented the responses of one subject. No claim is made that she is absolutely typical or representative since reactions varied a good deal in different subjects. However we can say that this kind of sequence of responses is in no way unusual. The subject is female and age nineteen. The responses are to Card I. Inquiry is omitted.



## Test 1.

- 1 — Bat
- 2 — Air Shield
- 3 — Cliffs
- v 4 — A water fountain as in a park
- 5 — A Chinese tower, place for cars to drive under

## Test 2.

- 1 — Hands reaching out
- 2 — Top—an explosion with fragments flying
- v 3 — Small animals in center casting spell downward

## Test 3.

- 1 — Two large birds resembling people fighting
- 2 — Two tiny figures like elves right in center,  
each has an arm up over his head
- 3 — Egyptian or Chinese lady with full sleeves  
and tall hat in center

## Test 4.

- 1 — Two winged chiefs over a conference table
- v 2 — Masses of molten lava

In this example new responses appear on tests two, three and four which not only lend themselves to new content interpretations but were not suggested by anything which came before. This is absolutely characteristic and occurred in every one of the twenty-eight subjects.

*Discussion*

What are the implications of these findings? The fact that these subjects have been able to produce large numbers of additional responses is, we believe, of considerable significance. It suggests to us that the very great individual differences in Rorschach productivity which constitute one of the most notable of Rorschach phenomena, do not stem from basic differences in capacity but are dependent instead on motivation and set and are to a considerable extent subject to manipulation. This is a very large jump from our results, especially since our subjects were college students, a group noted for its high productivity. It would therefore be very nice to have this experiment repeated with other classes of subjects, including that recalcitrant group of psychoneurotic patients who ordinarily average around twelve responses. Such experiments are planned for the near future. How-

ever, we do not feel that the correctness of our conclusion is really dependent on the results of these additional studies. Neurotic subjects in a hospital setting will undoubtedly be less responsive to our wishes than were the college students, but we would be very reluctant to believe that really basic differences exist. To get more responses from the former one would simply have to develop an appropriate test setting in which compliance with the tester's instructions satisfies the subject's needs in some genuine way. We do not mean to assert that intelligence is not related to the capacity to produce large numbers of responses but we think that subjects within the normal range have a much larger Rorschach potential than is ordinarily sampled by a first performance. Individual differences in capacity may exist at a much higher level so that some people may be able to produce as many as three or four hundred responses before quality declines and they are unable to go on while others may be able to give only seventy-five to a hundred responses. Our point is not that differences in capacity do not exist but that the number of responses given on a first performance is a poor measure of capacity.

Since a first Rorschach administration ordinarily samples a good deal less than the total Rorschach potential we may ask whether it adequately represents the larger universe of responses. Our findings are not clear on this point. On the one hand they indicate a definite similarity between the first and the second, third and fourth performances. On the other hand there is a considerable amount of change and the basic shape of the psychograms tend to be quite variable. The size of the correlations when judged by the standards against which we ordinarily measure reliability are quite low. However, there is a question as to whether these standards are appropriate in this situation.

While we entertained the possibility that the correlations might be very high and would have regarded this as convincing evidence of one aspect of the reliability of the test, namely performer consistency, our own expectations were that they would not be high since we took the position that only some aspects of the subject's personality would be expressed on a first test and that others would be expressed on subsequent performances. Our findings indicate that this expectation was correct. Nevertheless the amount of stability is considerable and may be regarded as evidence that the Rorschach test does tap some functions of personality which express themselves in a consistent way.

The demonstration that subjects can produce substantial numbers of additional responses inevitably raises the question of why they ordinarily stop so far short of their capacity. One possibility which we seriously entertained was that subjects become psychically fatigued or satiated and are



actually unable to continue. Certainly some subjects act very much like brain injured patients in their inability to shift and reorganize the blot stimulus despite apparently strong efforts to do so. However, these appearances may be deceptive and further research is needed to determine whether actual incapacity or motivation and set are the crucial factors.

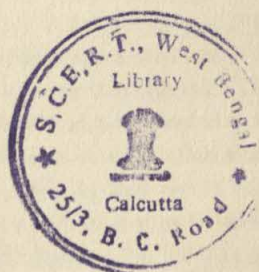
In conclusion, these results suggest that a certain degree of caution is appropriate in interpreting a single Rorschach performance. Such interpretations should be made in the light of the findings, presented in this study, that the responses obtained are only part of the story, and that they are tied to the peculiar circumstances under which the test was given. We interpret our data as demonstrating in conclusive fashion what many other Rorschach studies have implied; namely, that the single Rorschach performance cannot be regarded as an adequate, stable or complete representation of the personality characteristics which the Rorschach is able to describe.

## REFERENCES

1. EICHLER, R. M. Experimental Stress and Alleged Rorschach Indices of Anxiety. *J. Abn. Soc. Psychol.*, 1951, 46, 344-355.
2. GIBBY, R. S. The Stability of Certain Rorschach Variables Under Conditions of Experimentally Induced Sets: 1. The Intellectual Variables. *J. Proj. Techn.*, 1951, 15, 3-26.
3. HUTT, M. L., GIBBY, R., MILTON, E., AND POTTHARST, K. The Effect of Varied Experimental Sets on Rorschach Test Performance. *J. Proj. Techn.*, 1950, 14, 181-186.
4. KENDALL, M. S. *Rank Correlation Methods*. London: C. Griffin, 1948.
5. KIMBLE, G. A. Social Influences on Rorschach Records. *J. Abn. Soc. Psychol.*, 1945, 40, 89-93.
6. KLATSKIN, E. H. An Analysis of the Effect of the Test Situation upon the Rorschach Record: Formal Scoring Characteristics. *J. Proj. Techn.*, 1952, 16, 193-199.
7. LORD, E. Experimentally Induced Variations in Rorschach Performance. *Psychol. Monogr.*, 1950, 64, No. 10.

# VI

## *Current Status*





*L. L. Thurstone*

## THE RORSCHACH IN PSYCHOLOGICAL SCIENCE\*

THE RORSCHACH test has attracted so much attention since its introduction in this country about twenty years ago that it has come to occupy a unique position among the many hundreds of tests that are being used by the psychological profession. A few years ago young psychologists studied the Binet tests in order to be able to take jobs for the purpose of assigning IQ's to people. At the present time we do not hear much about such ambitions, but we do hear often from prospective students who want to study the Rorschach test in order to be able to qualify for jobs in giving this test. Comparison of these two situations gives an opportunity to call attention to some conditions of psychological service and research that need to be improved.

When the Binet tests were most in vogue, they were regarded as some sort of base criterion for judging all other tests. The Binet tests were regarded as if they constituted something so basic that all other work had to be oriented around them. There were hundreds of studies in which the investigator was proud to report that his test agreed well with the Binet test. There is a similar attitude at the present time with regard to the Rorschach test. It is regarded as if it were something scientifically unique. Whatever it

Reprinted from *J. Abnorm. Soc. Psychol.*, 1948, 43, 471-475, by permission of the American Psychological Association.

\* This paper was presented to the Illinois Association for Applied Psychology in Chicago, November 18, 1947.

is that makes the Binet test useful or whatever it is that makes the Rorschach test useful, we can be pretty sure that the same results can be obtained with different methods and with entirely different tests. We may not know what those other methods and tests are which could be used to explore the same domain as the Rorschach test, but it seems evident enough that those who are working most enthusiastically with the Rorschach ink blots are not making much effort to discover other entirely different methods of exploring the same domain. In this respect the situation is quite different from that of the Binet test popularity some years ago. Even then, there was a good deal of experimentation with a great variety of tests and comparisons of the results in terms of various kinds of appraisal of the subjects. There was serious study of the logic of the Binet tests and there was serious effort to relate it to the psychological concepts that were current. As a result, psychologists generally took part in deliberations about the underlying theory of the Binet tests. At present we have relatively little of such effort with regard to the Rorschach tests. Such discussion seems to be confined to a cultish group that has adopted its own jargon without relation to current experimental work and theoretical work in psychological science.

It would be fortunate if the Rorschach test could be removed from its isolation from the rest of the psychological profession. This interesting test gives results that occasionally demand our attention and it is a challenging problem to discover variant methods of getting at the same types of appraisal so that the nature of the underlying processes may be better understood. That is the scientific problem in which many other psychologists could participate even if they do not belong to the select few who have qualified as full-time Rorschach examiners. The first step in removing the Rorschach test from its isolation must be in translating the specialized jargon into currently known and accepted concepts, or else in the introduction of such new psychological concepts as may be necessary and with such discussion that psychologists generally can understand the new concepts. The result may be that psychologists will come to adopt some of the Rorschach terms in their own writing just as was the case some years ago when Freudian concepts were introduced into psychological thinking. But in order to bring about such a result, it is first necessary for the Rorschach students to make themselves understood among psychologists who may see no reason why they should study the specialized jargon that has been built up around a particular set of ink blots. The first task is for the Rorschach students to put their discussion in terms that can be understood by the psychological profession. The burden of proof is on them to show that they have not only a useful trick with a particular set of ink blots but that they also have some important ideas that psychologists should learn about. It is my own judg-



ment that the Rorschach test should be critically studied not only for the diagnostic value that it has shown in psychiatric work but also for the psychological principles and effects that are associated with projective procedures, of which the Rorschach is only one of many examples.

There is good reason why the projective methods have become very popular in psychological examining and there is also good reason for the projective methods to be incorporated more frequently into psychological experimentation. The projective method of examining consists essentially in giving to the subject an ambiguous presentation, an ambiguous task, which can be completed in many different ways. The theory of a projective test is that when the subject does respond to an ambiguous assignment, he is revealing himself in the response that he spontaneously makes. This is a fruitful device, as has been shown in a number of tests.

There is a common misunderstanding about the projective test method as regards the scoring of the performance. Many examiners seem to have the impression that a projective test cannot be objectively scored, but that is a mistake.

By way of example, I shall describe a projective test that is objectively scored. I assembled a list of about forty homonyms which had two common associations or meanings. Each one of the homonyms was so selected that one of the common meanings was human and social in significance while the other meaning was something physical or literal. The subject is asked to write as quickly as possible the meaning of each word, or an association which will indicate the meaning of the word. The responses are scored by counting the number of social associations. The individual differences are considerable and the exploratory experiments with this test indicate some relation to the temperament of the subject. Before suggesting that such a test be used for examining we would have to give it to people with known temperamental differences to determine whether it is trustworthy. My point here is merely to show one of many projective tests that can be objectively appraised if one has some psychological plan or idea in designing it.

In appraising a projective test performance, we may distinguish between two types of purposes. One purpose is to elicit from the subject some idea about his past history, the biographical detail that fits the particular case that is being examined. This can be done by free association, or by the Rorschach, or the Thematic Apperception Test, or by other projective methods. Biographical detail may be so emotionally blocked that it is more readily accessible by these indirect methods than by direct questioning. While this type of inquiry about a subject may be useful for some immediate purpose, we must remember that no amount of anecdotal or biographical detail will ever become science until someone organizes the individualistic material

into some categories of fruitful classification so as to reveal the underlying parameters of the dynamical system that constitute a personality. This is the second purpose for which a projective test may be given. It should reveal, not only the biographical detail that is of interest in a particular case, but it should also reveal the subject as to the dynamical characteristics that describe his motivations and values. Even more fundamentally, it should reveal the style of his personality so that his behavior becomes at least in some sense predictable as to the style of response that he is likely to give to different types of life situations. But this cannot be accomplished as long as we stay within the narrow confines of responses to the ink blots. These must be interpreted in terms of larger classes and types of response that transcend the ink blots. If such interpretations can be made in a dependable way, then they should be described in every textbook in psychology. The student reader should be shown the experimental evidence and the psychological theorizing by which the color shock, the movement responses, the animal responses, the responses to the white spaces, become significant in the larger setting of psychological interpretation.

On many occasions I have suggested to my friends who are working with the Rorschach that they should experiment with other projective test procedures in order to throw light on the underlying mechanisms and the response types of the subject that may be of interest in understanding his temperamental characteristics. Several kinds of projective procedures should be tried with the same groups of subjects, who should be appraised as to their temperamental characteristics as well as to their preferred types of adjustment to life problems. Instead of merely suggesting that more varied forms of projective methods be tried, I shall describe a few such tests.

Recently I have discussed with some of my friends a type of projective test for appraising personality or temperament. I do not know if it will be successful but I believe that it has good possibilities. Our plan is to make a collection of prints, perhaps fifty or one hundred. These should be so varied as to type of painting that practically everybody will find some prints that he rather likes and some that he definitely dislikes. The collection should be increased until we have such diversity that everyone likes at least some of the prints. Then we want to study the clustering so that, if a subject points to a few pictures that he likes, we should be able to predict some prints that he will like and some prints that he will dislike. The art preferences may be associated with temperament. For example, it is probable that preference for highly saturated color contrasts as against low color saturations may be significant. Preference for strong action as against passive or stationary objects, preference for lots of detail as against bold outlines, fine-line work as against coarse-line work, pictures with people and those without people,



photographic detail as against abstract design, and many other comparisons may be made in studying the clusters of pictures that are liked and disliked. These preferences may be related to personality. My suggestion would be to ask the subject to say nothing. I probably would ask him merely to indicate his likes and dislikes and I would have only a secondary interest in his explanations because they would probably be wrong. On the other hand, one might gain some insight from the subject's verbalizations about his preferences. One can tell by trying.

In this type of exploratory work, I should not care to have any test criteria beforehand. In this case we would depart entirely from the convention of correlating test scores with outside criteria. I would select two groups of subjects as regards their likes and dislikes, even if I had only a few people in each group. Then I would ask how these two small groups of people are different. If I could find some difference in temperament or style of personality, then I could make an hypothesis about what the difference in art preference might mean. I would proceed in the same manner for each type of preference or dislike.

The study of expressive movement is a field that should be actively cultivated experimentally by students of this problem. The early work of June Downey has been thrown into the discard by most psychologists because of criticisms about minor questions of reliability and because of validity studies, mostly with the wrong criteria. It is my judgment that she was on the right track in studying personality by objective experimental methods. The studies of Gordon Allport and others on expressive movement should be continued in the hope of finding objective experimental methods for appraising personality. One could easily list some forty or fifty experimental studies that should be made in the direction of studying personality by objective and experimental methods.

It would be fortunate if students of the Rorschach test would proceed to qualify themselves for membership in the psychological profession by insisting on experimental evidence under reasonably controlled conditions for the various interpretations that they make of the responses to a set of ink blots. There is justification for the belief that the projective test method of presenting to the subject an ambiguous task has great possibilities, and it seems plausible that a set of ink blots may be one of these fruitful test methods. Rorschach students are not making any worthwhile contribution to psychological science as long as they remain in their state of gullible and uncritical acceptance of fanciful interpretations of the responses to a single set of ink blots. If they don't have imagination to do anything more challenging, they can at least try, say, twenty different variations of the ink blots. The designs can be varied in the relative degree of freedom of percep-

tual closure for the subject, and in many other ways. The various types of ambiguous presentations should be studied experimentally and systematically on the same subjects in order to extract the psychological principles in this domain. When these principles begin to be understood, it will probably become apparent how the test methods may be still further improved. This progress will not be made as long as we stick to a single set of ink blots without relating them experimentally to the many variations of the projective method.

One Rorschach student has said that all he needed to know was that the psychiatrist wanted the Rorschach test. That justified the Rorschach test. There are two possible interpretations of such a remark. It may mean that the psychiatrist has so little technique himself that he grabs at anything, even a Rorschach. Another interpretation is that the psychiatrist thinks that this is all he can get out of a psychologist. Maybe he is right.

When I was asked to participate in this program, I assumed that it was not because of qualification as a Rorschach examiner. While I have included it in some of my studies, I have not been a student of the Rorschach.<sup>1</sup> For many years I have been associated with the development of psychological tests of considerable variety and purpose and with the development of test theory. I have been interested in the projective method which we have used in a number of experiments. My purpose in accepting participation in this program was not to elaborate on the possibilities and limitations of a particular set of ink blots. Rather, I have attempted to show why it is that the Rorschach test has not been accepted by the psychological profession, and why it is that most students of this test do not have recognition or status in psychological science. I have attempted to outline how this condition can be improved and how the students of the Rorschach might investigate their problems so that this field of inquiry can be incorporated into psychological science.

1. At the time when the Rorschach test was introduced to American psychologists, I discussed with Dr. David Levy the possibility of objectifying the scoring of the Rorschach test. It was my judgment that the appraisal of a Rorschach performance could not profitably be objectified without much experimental work.



*Joseph Zubin*

## *FAILURES OF THE RORSCHACH TECHNIQUE*

**I**F HERMAN RORSCHACH had not been carried off by the infection following a minor operation, he would have celebrated his sixty-eighth birthday last November 8. He would have looked approvingly on our efforts and would have helped make this symposium on failures a success by providing methods for their elimination in the future. For no one was more tentative in his conclusions, more demanding of continual self-survey and revision than the man who lent his name to the major projective technique. "The conclusions drawn," he says, "are to be regarded more as observations (remarks) than as theoretical deductions. The theoretical foundation for the experiment is, for the most part, still incomplete." He even went as far as hoping for "... control experiments taking up each symptom individually, and other psychological methods which might also be used in control research" (24).

It is both a brave and a wise move that the Society for Projective Techniques has undertaken in this symposium and it is to be congratulated on its maturity and integrity. Like the post-mortem in surgery, failures in projective techniques ought to be far more revealing and instructive than the reports of successful cases. We stand to benefit from our mistakes, find out what changes need to be introduced, what hypotheses need to be altered

Reprinted from *J. Proj. Tech.*, 1954, 3, 303-315, by permission of the publisher and the author.

and what expectations need to be amended. Following the custom of the surgeons, only the outstanding failures of the Rorschach technique will be discussed, the outstanding successes being omitted. A cataloguing of one's failures without scanning the successes is trying to even the most mature of souls, but since it is motivated by a search for improvement we can all bear up under it. After analyzing the failures, a hypothesis to explain them will be presented, which may perhaps repay, in part, for the masochistic trend a self-analysis may engender.

One outstanding Rorschach worker has made the following surmise about Rorschach's own reaction to the current scene: "Rorschach today would still recognize the cards. Brilliant though he was, I doubt if he could find time to read the voluminous literature which is well en route to the thousand mark. I am certain that, if he could, he would be startled that from his little experiment following the 'scoring' of the responses, there emerge invaluable facts relating specifically to the way in which the patient sees his world, approaches and handles it, and of what this world consists. His anxieties and insecurities, his hurts and wishes, his fictions, his needs, his assets and liabilities, his likes and dislikes—all of these and more emerge to be viewed by the examiner. Moreover, the pattern reveals also the meaning of these things to him, the configuration of his personality which thus results, and the motivations of his behavior. It, furthermore, aids in differential diagnosis, particularly between the organic and functional types of illness, and among the affect and content disorders. The expert examiner can also obtain from the response record a practical estimation of such important personality features as intellectual efficiency, emotional maturity and balance, and degree and depth of reality acceptance. Finally, the procedure serves as a guide to therapy and an index of its success or failure" (19).

What validity do these claims have? Let us take a look at the record. The following questions need to be answered: 1) What reliability does the Rorschach technique possess? 2) What validity? 3) What relationship does it bear to the changes brought on by therapy, and 4) to the outcome of therapy, and 5) what light has recent research cast on these problems?

First, what about the reliability of scoring? This question has received but little attention and it is generally taken for granted that scoring has a high degree of reliability. Hertz (18), however, has stated on the basis of her long experience that "scoring still remains a matter of skill—art, if you will." Though this statement was made eighteen years ago, it still largely holds true today. Ramzy and Pickard (22) found that only after considerable discussion and arbitrary acceptance of certain conventions were they able to obtain consistency in their scoring. It is noteworthy that appeal to textbooks only increased their confusion. Even after this collusion, the degree of agree-



ment was only 90 per cent for the location category. Since Beck's location tabulations were followed, it is surprising that the degree of agreement was not perfect. For determinants of Form and the Movement variety, the agreement dropped to 83 per cent, and for the Color and Shading determinants to 75 per cent. As far as content was concerned, following Beck's classifications an agreement of 99 per cent was obtained. Ames and her associates report their reliabilities in terms of product moment correlation coefficients: location categories, .92; form and movement categories, .90; color and shading, etc., .80, and content categories, .97 (1). These results are essentially in keeping with those reported by Ramzy and Pickard. Since according to Rorschach, "The actual content of the interpretations comes into consideration only secondarily" the determinants being of major importance, it is clear that the best estimate of agreement for the major scoring categories is only 80 per cent. This is a far cry from the degree of agreement expected of an individual test. The lack of objectivity in scoring is further evidenced by Baughman's experiment (5) in which fifteen Veterans Administration examiners disagreed significantly in scoring on sixteen out of twenty-two scoring categories in records based on a random selection of cases culled from their files. It is clear that failure to provide an objective scoring system is our first failure.

The reliability of the test can not be judged by split-half methods because of the heterogeneity of the blots, nor can it be judged by test-retest because of the memory factor. Alternate forms are required for this purpose. Eichler (12) reports that he tried to obtain a reliability estimate by correlating the Behn-Rorschach with the Rorschach but obtained such low reliabilities that he concluded the two forms were similar but not parallel. Thus, lack of reliability is the second failure.

The validity of the test will be analyzed from the following points of view: (1) subjective, (2) clinical, (3) statistical, and (4) experimental.

### *Subjective and Clinical*

Subjective validation of the testimonial variety in which he who comes to scoff remains to pray, will not be commented on further. This type of evidence is so clearly discountable as utterly unscientific that it is not even to be counted among our failures.

The clinical method sometimes consists of administering and scoring the test, collecting data on the subjects' subsequent behavior and then going back to the protocol, in which are "found" signs which "unmistakably" foretell such behavior. Unless a cross validation of these signs is undertaken in

another study, it is fruitless to accept them as indicative of future behavior, because with sufficient imagination and exertion of effort through trial and error, pseudosignificant signs can be found in any test. Unfortunately, such cross validation is rarely encountered.

Blind analysis is one of the spectacular aspects of the Rorschach technique and has probably been the most important factor in the acceptance of the Rorschach. One would wish that this method could be made more explicit and more public, and that the enthusiastic proponents of this method were as ready to publish their failures as their successes. Until this method becomes more open to public scrutiny, it has to be placed in the doubtful category and counted neither as a success nor as a failure.

The matching technique is another way of demonstrating validity. Unfortunately, there are many inadvertent and tangential characteristics in this method, not germane to validity, which may influence the outcome. Successful matching is frequently effected on the basis of minor details or coincidences, rather than essential equivalence. Heterogeneity of matches also makes the task too easy. Determination of the precise ground on which successful pairing is made is virtually impossible. Furthermore, most of the results indicate only that the matching is better than chance, an insufficient criterion for validity. Cronbach (9) has recently devised a trenchant methodology for freeing the matching methods from these defects, but it is quite intricate, and the one application which he made yielded no success in the matching process. Thus, the clinical evidence for validity cannot be accepted scientifically, even though it is impressive. Our failure to provide more cogent evidence for clinical validity must be regarded as our third failure.

### *Statistical Studies*

In view of Rorschach's original purpose of devising a diagnostic test for mental disorders—especially for schizophrenia—we shall first review studies that deal with the diagnostic efficacy of the test. There are very few studies in which clinical scoring and interpretation disagree markedly with clinical diagnosis when the study is conducted in the same clinic and when the two diagnosticians have had considerable experience in working together. This is certainly highly in favor of the test, and the only question a carping critic might ask is: is there a tendency for collusion to occur in such cases, since only a few authors, such as Benjamin and Ebaugh (4), point out the care they took to avoid collusion? When we examine the relationship between the individual Rorschach scores and diagnosis, a totally different picture emerges. Guilford (16) found in three successive samples of about fifty



neurotic patients, who were given the Rorschach in orthodox fashion, that no significant differences could be detected between their performance and that of a large normative group of cadets. Wittenborn and Holzberg (32) found zero correlation between Rorschach factors and diagnosis in 199 successive admissions. Cox (8) found only five scores out of a total of forty-three scores differential between normal and neurotic children, and of these five, three were in the content categories and only two in the determinant categories. These are only samples of the well-known failure of the individual Rorschach scoring categories to relate to diagnosis, which is in marked contrast to the success of the global evaluation claimed by clinical workers generally. This must be regarded as the fourth failure.

In prognosis too, the recent review by Windle (30) leaves one with very little faith in the efficacy of the prediction attributable to the Rorschach. The only successful prognostic elements seem to be based on content rather than formal factors, as shown by McCall (21). This is the fifth failure.

Since Windle's article was published, two additional bits of evidence of failure of the Rorschach in the prognostic sphere have appeared. In Barron's study (2) in which the Rorschach, together with several other tests including the MMPI were given to both patients and therapists before the beginning of therapy, while the MMPI predicted outcome significantly, the Rorschach, despite all attempts ranging from the global to the atomic, failed to do so. Rogers, Knauss, and Hammond (23) report a similar experience. As long as other tests failed to predict outcome one might have attributed the failure to the heterogeneous nature of the patient group—to an admixture of early and chronic cases, for example. When other tests succeed where the Rorschach fails, one can either conclude that the Rorschach is unsuited to prediction, or that basic personality which the Rorschach claims to measure is unrelated to the type of therapy involved.

The differentiation of organic from functional conditions has also had a checkered history of success and failure: success when viewed retrospectively but failure when the results of retrospective analysis were applied to a new sample. The latest in the series is the study by Dörken and Kral (11), who after demolishing the signs of previous workers propose a new set of their own, which will probably in turn be demolished by the next worker. The vitality of this search for signs despite so many failures can only command the awe and respect of the onlooker. Whether it will finally succeed only time can tell; meanwhile it must be placed in the doubtful category.

The success that clinicians have had in global evaluation of mental patients has not been duplicated in the global evaluations of normals. Here the series of failures is truly appalling. The story is too long to review. Some of the recent examples are—Grant, Ives, and Ranzoni (15), who found zero

correlation between Rorschach evaluations and case history evaluations of adjustments in eighteen-year-old normals. The more specific adjustment signs provided by Helen Davidson (10), fared a little better, yielding correlations from .23 to .56. The failure of the Rorschach to serve as a predictor of success in the screening programs of the armed forces, in the screening of clinical psychology students and students of psychiatry, is too well known to warrant further comment. I shall quote from only one of these:

It was regarded as very important that the Rorschach test should be given full opportunity to show what it had to offer in a personnel-selection setting. It was recognized that neither time nor personnel requirements for the routine administration and use of this test were consistent with the mass testing required. . . . Yet the test was administered experimentally to several hundred students individually according to the prescribed procedures by members of the Rorschach Institute who were serving in one of the psychological units. Two methods of group administration were also tried, the Harrower-Erickson and our own version.

The results were almost entirely negative. From the individual administration of the test, neither the 25 indicators taken separately or collectively nor the intuitive prediction of the examiner based upon the data he had from the administration of the test gave significant indications of validity against the pass-fail criterion. There were two samples, one of nearly 300 and the other of nearly 200. The Harrower-Erickson group-administration form also gave no evidence of being valid for pilot selection. The AAF group-administration form when scored for the number of most popular responses showed a coefficient of .24, based upon a sample of more than 600 students (16).

The inability to differentiate between normals is the sixth failure.

Special studies aimed at evaluating intelligence by means of the Rorschach also usually come a cropper. Wittenborn (33) compared the extreme groups selected on the basis of college entrance examinations on eighteen scores of the Rorschach and failed to find any significant relationship. (This is even worse than chance, because by chance you might have expected about one of these eighteen scores to show significance on the .05 level.) The relationship of Movement to intelligence has been investigated by Tucker (28), in 100 neurotics, who found a very low correlation—.26. Wilson (31) made a more extensive study of a large college population and used Movement, Form level, Whole, Responses, Z (organization), diversity of content, and a new specially-designed-variable designated as "specification"—and found zero correlation with intelligence.

The studies in creative ability conducted in our own laboratory on creative versus non-creative writers, mathematical statisticians, and high school students have failed to reveal any differences on Rorschach performance and even tests especially designed to elicit Movement have failed (34). This is the seventh failure.

But the story of the use of the Rorschach with normals is not entirely a



hopeless one. Sen (26), in England, one of Cyril Burt's students, applied the Rorschach to 100 Indian students who had lived together for at least two years. Scoring by means of Beck's scoring system, the correlations with personality evaluations by their colleagues were non-significant. However, when scored for content *à la* Burt, the correlations ranged from .57 to .66. When matching was resorted to, a global method, both scoring methods yielded a high degree of success: .85 for Beck's system, and even higher for Burt's system. Interestingly enough, however, when a factor analysis was performed on the Beck Scores and on the Burt Scores, the results of both analysis are equally trenchant in their relationship between the derived factor scores and personality. This is a general finding in the studies in which the raw Rorschach scores failed to relate to diagnosis or personality. When factor analysis is resorted to, a rotation of the factors usually permits certain striking correlations with behavior to emerge.

We might stop here a moment to differentiate between content analysis as used by Burt and in our own laboratory and the content category as used by Rorschach. In the Rorschach there are really three types of classifications: Location, Determinants, and Content. The Location and Determinant categories are usually spoken of as the Formal Categories, and the Content Categories are those which simply classify the percepts as animal, vegetable or mineral—so to speak—that is, the category of objects it belongs to. It might be better to contrast in the Rorschach the perceptual factors—or structural factors with the content categories. There are left reaction time, popular responses, and confabulations, contaminations, etc.—which are neither determinants nor locations and hence could be classified with content. It is the non-perceptual part of the Rorschach performance—the thought content—which we designate as the content aspects of Rorschach performance.

Examples of the scales used for content analysis of the Rorschach protocols are: 1) Formal content *à la* Rorschach, 2) Dynamic content—(a) degree of evaluation included in response as judged by qualifying adjectives, (b) degree of dehumanization, (c) ascendance-submission in concepts portrayed (slaves-versus-kings, for example), (d) definiteness of concept (e) abstractness (f) dynamic qualities—alive or dead, static or moving, (g) distance in time and space (h) self-reference, (i) perseveration, (j) elaboration, (k) blot versus concept dominance, (l) interpretive attitude, etc. (33). Further evidence for the success of this type of analysis is found in Elizur (13) and in Watkins and Stauffacher (29) in this country and Sandler (25) in England.

Elizur found that an analysis of content in relation to hostility yielded significant correlations with ratings of hostility. Sandler, working with Rorschach's content categories (and not with the type of content analysis being discussed here) made a factor analysis of the content scores of fifty psychi-

atric patients at Maudsley Hospital, ranging over eight types of mental disorder. He emerged with four factors and determined the psychological meaning of each factor by its correlation with the personality evaluation made by psychiatric interview and case history methods. These were drawn from three levels—previous personality, general background data, and present symptoms. The productivity factor, *R*, for example, was highly related to previous productivity in life, to chronicity of symptoms and to a schizo-affective picture at time of hospitalization. The Anatomy factor—internal anatomical objects versus external objects as another example, was related to an insecure, withdrawn, “previous” personality picture, bad physical health and an emotional deluded state for the “present symptoms.” The remaining factors were analyzed in similar fashion.

Watkins and Stuffacher (29) provided a series of indexes of “deviant verbalizations” based on the content of the protocols and found that such indicators had a reliability of .77 between two raters, and that these indexes distinguished normals from neurotics and the latter from psychotics.

Factor analysis, when applied to either the orthodox scoring categories or to the content scales, emerges with factors like the following: fluency, generalizing ability, emotionality, imagination, extraversion-introversion, neurotic tendencies. Apparently, what the Rorschach expert does intuitively in evaluating the records of normals and neurotics can be obtained objectively by factor analysis. But it should be noted that direct statistical manipulation of the original Rorschach scoring categories does not lead to significant results unless they are distilled either through the mind of the expert or the hopper of factor analysis.

As for the relation of the Rorschach to changes accompanying psychotherapy, the results are in doubt. One study claims positive findings (21a) and three show negative results—Lord (20), Carr (6), and Hamlin and Albee (17). The latter found that Muench's indicators of improvement did not hold up when groups exhibiting different levels of adjustment were compared, thus negating the one positive study mentioned.

Barry, Blyth, and Albrecht (3) compared test and retest data on the Rorschach with pooled judgment of patients at a Veterans Administration Mental Hygiene Clinic. Changes in ratings of adjustment level failed to correlate with changes on the Rorschach.

The recent topectomy study (7) offered an opportunity for testing what effect the lowering of anxiety induced by the operation might have on Rorschach performance. Neither orthodox scoring nor anxiety indicators (with the single exception of reaction time) succeeded in demonstrating any changes in the Rorschach performance of the patients, although other psychological tests showed such changes. Psychometric scaling, however, did re-



veal certain changes and also provided a prognostic indicator. Three pairs of patients were selected, each pair consisting of one individual who decreased in anxiety and one who increased in anxiety after operation. The judgment of loss and gain in anxiety was based on psychological interviewing by means of anchored scaling devices and on the judgment of the psychiatrist. Only those patients in whom the two criteria concurred were selected. The results indicated that perception of movement of whatever variety, regardless of whether it was accompanied by empathy, correlated positively with anxiety, rising when the anxiety level rose and dropping when the anxiety level fell. The degree of tentativeness or insecurity in giving responses also correlated positively with anxiety. The following variables showed only a unilateral relationship to anxiety levels, declining with a decline in anxiety but showing no corresponding rise with rise in anxiety level: sensitivity to chiaroscuro, anatomical responses, perception of animate objects, perception of objects with texture, and degree of self-reference. The following variables also showed a unilateral but negative relationship with anxiety, showing increases as anxiety fell—accuracy of form perception and degree of congruity of the response. The statistical significance of these differences could be readily established since each patient could be analyzed as a separate sample and the significance of the difference for each patient determined. Only the variables that showed consistent changes from patient to patient were reported.

The fact that the classical Rorschach scoring is not sensitive to changes induced by somatotherapy is an old story. Lord (20) reports a similar finding in psychotherapy. Perhaps the Rorschach test reflects only basic personality structure. Someone has suggested that the goal of therapy is to arrest diseases into defects and then teach the patient to accept these defects. If therapy consists in nothing more than the acceptance of one's disabilities, no change in fundamental personality is to be expected.

### *Experimental Studies*

Perhaps the most important question that the experimentalist would like to answer about the Rorschach technique is: what is the stimulus, what role does it play, and whether present scoring of stimulus qualities such as Color, Form, Shading and perhaps Movement have definite stimulus correlates. One way of answering this question is to alter the stimulus characteristic to see whether the responses will change correspondingly.

The most revealing study of the stimulus properties of the Rorschach is a still unpublished study by Baughman (5). He set as his goal to differen-

tiate as far as possible between that part of the response which inheres in the stimulus and that which inheres in the responder himself. Since the characteristic of the stimulus are more readily manipulated, he devised a series of modifications of the Rorschach plates so as to reveal the potency of a given part of the stimulus for evoking characteristic responses. He started off with the standard card and eliminated first the hue factor through photographing the standard series in black and white on panchromatic film, retaining all the nuances of the shading or differences in brightness and all the other characteristics of the blot. Then, he removed the shading by making line drawings of the more striking contours within the blot and the periphery. The third modification consisted of blotting out the inside details by making the entire inside of the blot black but leaving the islands of pure white, yielding a silhouette effect. The final modification consisted simply of the periphery or outline of the blot.

While it is difficult to draw up a correspondence between the altered appearance of the card and the specific Rorschach determinant which is most prominently present or absent, the following tentative suggestions can be made. The cards in which only the periphery was present would tend to accentuate whole responses and form responses. The cards with the inside details would tend to accentuate detail responses and perhaps organization, (Beck's Z). The silhouette cards would tend to accentuate form responses and perhaps tend to suppress white space responses. The achromatic cards and the complete original set are too well known to require further discussion. The modified cards as well as the original set were administered to a group of 100 veterans, hospitalized for neuroses and character disorders. Each of the five series of blots was given to a group of twenty patients selected randomly which was equated with the other groups on IQ and educational level. Beck's scoring system was employed.

When he compared the records of the various groups, which had been administered variants of the original stimulus cards, he found, instead of the theoretically expected changes, a considerable degree of constancy in the responses. Virtually all of the significant differences were attributable to differences in stimuli which were simply objectively necessary for the occurrence of a given category of responses. For example, Detail responses were found uniformly distributed in all the variant forms of the blots except for a drop in the case of the peripheral form series, when the stimulus for D is eliminated. Apparently color and shading are not important for the Detail category. Surprisingly, the M response occurred with significantly higher frequency in the silhouette version, indicating that perception of movement is independent of shading. Categories showing but slight differences from series to series were: R, W, Form level, P, T/R, A, FM, Total time, Diversity



of content, and 8-9-10%. Baughman aptly summarizes these findings: "The severest assault upon the stimulus is necessary before significant changes in resulting performances are produced."

Another analysis of the data was made by submitting the protocols, including the reaction time data and the responses, to experienced Rorschach workers to see whether color and shading shock patterns occurred only in the appropriate series. As a result of this investigation, it is reported quite conclusively that the time latency and response patterns supposedly typical of color shock occur with the same frequency whether color is present or absent. The same was found to hold true of shading and shading shock. A very substantial question is thus raised as to the wisdom of continued use of the shock indicators. Further evidence against the concept of "color shock" is provided by Siipola (27). In an especially ingenious experiment, she concluded that the affinity that color bears to emotion is based on a misunderstanding. Color shock for example is not due to color itself, but to the incongruity between the color and contour of a given blot area. The conflict engendered in the observer will take different paths depending upon the personality of the subject. While Siipola's experiment is not itself conclusive, no statistical verification being given, her explanation of color shock is ingenious, to say the least.

Many of the moot questions in scoring could be answered by Baughman's research. Thus, "Bat" to Card I practically disappears as a response when "shading" and "black" are removed, while "Butterfly" occurs equally frequently in all the modified presentations. Similarly, "Map" occurs only when shading is present. Colored areas which yield anatomy responses practically cease to do so when color is eliminated. Color is of little importance in the response "Bat" to Card II—D32, in "Monkeys" in Card III—D3 and in "Bow tie" and "Ribbon" to the same card. "Rejection" in this study was much more prominent when Color and Shading were absent, indicating that Form rather than Color or Shading is the primary source of rejection. This study needs to be repeated on groups other than the neurotic, with other scoring systems, and with other types of modification.

The second question deals with the effect of alteration of the state of the subject by means of drugs or hypnosis or shock. These experiments have not yielded very much because of the inexactness with which the psychological correlates of these induced states are known.

The third question deals with the effect of alteration of the circumstances surrounding the test. Prestige suggestions as to the importance of certain types of responses will alter the distribution of the responses in the direction of the prestige. Similarly, social situation, induced anxiety, etc., have been tried out. Some of the changes expected by Rorschach workers

were validated, others not. The important conclusion to be drawn is that standard conditions are required and that Rorschach performance is not as insensitive to external conditions as some workers have claimed.

To summarize our findings thus far, the following facts are seen to emerge from our survey:

- (1) Rorschach scoring and sign evaluation has an *a priori* basis which is not always validated by experimentally contrived techniques such as alteration of stimulus, alteration of state of the organism, etc.
- (2) Globally, the Rorschach is an apparent success when the Rorschach diagnostician and the clinical diagnostician work closely together.
- (3) Atomistically it is an apparent failure.
- (4) Content evaluation whether done globally or atomistically is a success.
- (5) Factor analysis of atomistic scores consisting of the usual combination of perceptual and content factors, or of content alone, correlate with personality.

What kind of a hypothesis, what kind of a model, could satisfy the above conditions? That is the scientific question before us. Before answering, let us examine Rorschach's technique as an experiment. Experiments must have as their minimal requirements—subject, experimenter, apparatus or stimulus of some kind, a well-defined task, directions for the task, acceptance by the subject of the task, a response made by the subject and recorded automatically or by the experimenter. But that is not all—the most important part is still missing—the hypothesis.

What is the hypothesis underlying the Rorschach experiment? Rorschach never stated it explicitly, but it can be stated as follows:

- (1) We perceive in the artificial Rorschach space in the same way we perceive in real space.
- (2) The way we perceive in real space is determined by our personality.

Both of these assumptions are impossible to test at this time because we do not know how perception takes place in real space, nor how it takes place in Rorschach space. Gibson (14) has laid the foundations for the experimental determination of these two processes but we still have a long way to go before we can experiment with them. The relation between perception and personality must await the solution of the first two problems. What can be done meantime, and how can we explain the five facts which I have listed previously?

One hypothesis that suggests itself and which I humbly think merits consideration requires a shift of emphasis from the perceptual to the content



aspects of the Rorschach. It is true that Rorschach veered away from the content analysis of ink blots which was so popular with the psychologists of his day and espoused the formal aspects. He states "The content of the interpretations . . . offers little indication as to the content of the psyche." But he may have been wrong, or may have defined content too narrowly. If we define content as the essential elements of the protocol, and regard it as one would regard any other interview material, and analyze the content, the mystery is solved. Once the perceptual scoring is eliminated, and instead a content analysis of the verbal productions of the subject is made according to such categories as compulsive thinking, disorganized thinking, or creative thinking; poverty of ideas or fluency; confabulation or clarity; rigidity or flexibility; contamination or its opposite; perplexity or straightforwardness; rejection or compliance, etc., it will be discovered that such characteristics reveal themselves in the Rorschach the way they reveal themselves in the psychiatric interview. To be sure, the Rorschach interview is a standard interview and may lead to results which the free psychiatric interview can not lead to. But it is still an interview—an interview behind the veil of ink blots.

This would explain why content of Rorschach protocols is related to personality, whether evaluated globally or in isolated scales, while formal Rorschach factors fail to relate to personality. This would also explain why factor analysis of both formal as well as content factors relate to personality. In the course of the analysis, the content factors affecting the formal scores are teased out—viz.—the kind of mental content which serves to reduce R, disorganize F, disembody C or Sh, or prevent good M from arising in the mental patient and, *mutatis mutandis*, the kind of mental content which increases productivity and good responses in the normal, reveal themselves in the rotated factors. If this hypothesis be true, we should turn away from the indirect expression of mental content through determinants and location, and begin building scales for analyzing the content of the verbal productions directly. Such a beginning has been made by several workers and if we spend but 10 per cent of the harnessed energy behind the Rorschach wheel to studying the interview basis of the Rorschach, we may bring nearer the day when the contradictions that now exist within the Rorschach field are resolved.

New developments in the interview itself are fast turning it into a scientific tool, and since the interview, in the last analysis, is the basis for personality evaluation, no test today can rise above it. If we obtain objective criteria via the interview for the classification and evaluation of personality, perhaps such criteria may serve as a basis for the validation of tests.

But without an anchored interview, we float aimlessly in the sea of personality without compass or rudder.

Summary: This review of the failures of the Rorschach technique has found the following outstanding relationships:

- (1) Global evaluations of the Rorschach seem to work when the Rorschach worker and the clinician work closely together.
- (2) Atomistic evaluation, as well as global, of the content of the Rorschach protocols (as distinct from the perceptual scoring) seem to work.
- (3) Atomistic analysis of the perceptual factors is a failure.
- (4) Factor analysis of atomistic scores of both the perceptual as well as the content variety, seem to work.

The best hypothesis to explain these four facts is that the Rorschach is an interview and that its correct evaluation, like the correct evaluation of any interview, is dependent upon its content. If we provide scales for analyzing its content, we shall be well on the way toward clarifying many of the present day contradictions and obtain a better perspective on the evaluation of personality.

## REFERENCES

1. AMES, L. B., LEARNED, J., METRAUX, R. W. AND WALKER, R. N., *Child Rorschach Responses*. New York: Paul B. Hoeber, Inc., 1952.
2. BARRON, F. X. Psychotherapy as a special case of personal interaction: Prediction of its course. (Doctoral Thesis, University of California, Berkeley, 1950) quoted in Sanford, N. *Psychotherapy*, Stone, C. P., Editor, *Annual Review of Psychology*, Annual Reviews, Inc. Stanford, Calif., 1953, p. 338.
3. BARRY, J. R., BLYTH, D. D. AND ALBRECHT, R. Relationship between Rorschach scores and adjustment level. *J. Consult. Psychol.*, 1952, 16, 30-36.
4. BENJAMIN, J. D. AND EBAUGH, F. G. The diagnostic validity of the Rorschach test. *Amer. J. Psychiat.*, 1938, 94, 1163-1178.
5. BAUGHMAN, E. E. Rorschach scores as a function of examiner difference. *J. Proj. Tech.*, 1951, 15, 243-249.
- 5a. ———. *A comparative study of Rorschach forms with altered stimulus characteristics*. Ph.D. dissertation, Chicago, Illinois, March, 1951.
6. CARR, A. C. Evaluation of nine psychotherapy cases by the Rorschach. *J. Consult. Psychol.*, 1949, 13, 196-205.
7. COLUMBIA GREYSTONE ASSOCIATES, F. A. METTLER, Editor. *Selective Partial Ablations of the Frontal Cortex*, New York: Paul B. Hoeber, 1949.
8. COX, S. M. A factorial study of the Rorschach responses of normal and mal-adjusted boys. *J. Genet. Psychol.* 1951, 79, 95-115.



9. CRONBACH, L. J. A validation design for qualitative studies of personality. *J. Consult. Psychol.*, 1948, 12, 365-374.
10. DAVIDSON, H. *Personality and economic background: a study of highly intelligent children*. New York: Kings Crown Press, 1943.
11. DÖRKEN, H. J. AND KRAL, A. The psychological differentiation of organic brain lesions and their localization by means of the Rorschach test. *Amer. J. Psychiat.*, 1952, 108, 764-770.
12. EICHLER, R. M. A comparison of the Rorschach and Behn-Rorschach ink blot tests. *J. Consult. Psychol.*, 1951, 15, 185-189.
13. ELIZUR, A. Content analysis of the Rorschach with regard to anxiety and hostility. *Rorschach Res. Exch.*, 1949, 13, 274-284.
14. GIBSON, J. J. *The perception of the visual world*. Boston: Houghton Mifflin, 1950.
15. GRANT, M. Q., IVES, V. AND RANZONI, J. H. Reliability and validity of judges' ratings of adjusting on the Rorschach. *Psychol. Monogr.*, 1952, 66, 1-20.
16. GUILFORD, J. P. Some lessons from aviation psychology. *Amer. J. Psychol.*, 1948, 3, 3-11.
17. HAMLIN, R. M. AND ALBEE, G. W. Muench's tests before and after non-directive therapy: a control group for his subjects. *J. Consult. Psychol.*, 1948, 12, 412-416.
18. HERTZ, M. R. The Rorschach ink blot test: historical summary. *Psychol. Bull.*, 1935, 32, 33-56.
19. KELLEY, D. M. Clinical reality and projective techniques. *Amer. J. Psychiat.*, 1951, 107, 753-757.
20. LORD, E. Two sets of Rorschach records obtained before and after brief psychotherapy. *J. Consult. Psychol.*, 1950, 14, 134-139.
21. MCCALL, R. J. *Psychometric records in brain-operated patients*. Unpublished Ph.D. dissertation. Columbia University, 1950.
- 21a. MUENCH, G. A. An evaluation of nondirective psychotherapy by means of the Rorschach and other tests. *Appl. Psychol. Monogr.*, 1947, No. 13.
22. RAMZY, I. AND PICKARD, P. M. A study in the reliability of scoring the Rorschach ink blot test. *J. Gen. Psychol.*, 1949, 40, 3-10.
23. ROGERS, L. S., KNAUSS, J. AND HAMMOND, K. R. Predicting continuation in the therapy by means of the Rorschach test. *J. Consult. Psychol.*, 1951, 15, 368-371.
24. RORSCHACH, H. *Psychodiagnostics* (translation by P. Lemkau and B. Kronenburg). Berne: Verlag Hans Huber, 1942.
25. SANDLER, J. AND ACKNER, B. Rorschach content analysis: an experimental investigation. *Brit. J. Med. Psychol.*, 1951, 24, 180-201.
26. SEN, A. A. A statistical study of the Rorschach test. *Brit. J. Psychol.*, 1950, 3, 21-39.
27. SIIPOLA, E., KUHN, F. AND TAYLOR, V. Measurement of the individual's reactions to color in ink blots. *J. Pers.*, 1950, 19, 153-171.
28. TUCKER, J. E. Rorschach human movement and other movement responses in relation to intelligence. *J. Consult. Psychol.*, 1950, 14, 283-286.

29. WATKINS, J. G. AND STAUFFACHER, J. C. An index of pathological thinking in the Rorschach. *J. Prog. Tech.*, 1952, 16, 276-286.
30. WINDLE, C. Psychological tests in psychopathological prognosis. *Psychol. Bull.*, 1952, 49, 461-482.
31. WILSON, G. P. *Intellectual indicators in the Rorschach test*. Unpublished Ph.D. dissertation. University of Texas, 1952.
32. WITTENBORN, J. R. AND HOLZBERG, J. D. The Rorschach and descriptive diagnosis. *J. Consult. Psychol.*, 1951, 15, 460-463.
33. WITTENBORN, J. R. Certain Rorschach response categories and mental abilities. *J. Appl. Psychol.*, 1949, 33, 330-338.
34. ZUBIN, J. *Experimental abnormal psychology*, 1953 (mimeographed edition).



Lee J. Cronbach

STATISTICAL METHODS  
APPLIED TO RORSCHACH  
SCORES: A Review<sup>1</sup>

WHILE THE Rorschach test grew out of clinical investigations, and is still primarily a method of individual diagnosis, there is increasing emphasis on statistical studies of groups of cases. On the whole, the statistical methods employed have been conventional, even though the Rorschach test departs in many ways from usual test methodology. The present review proposes to examine the methods which have been employed to deal with Rorschach data, and to evaluate the adequacy of those often used. It attempts to provide a guide to future investigations by indicating statistically correct studies which can serve as models. There is no intent here to review the generalizations about the test arising from these studies, or to call into question general research procedures, sampling, and other aspects of the studies.

This report may be considered an extension of a review by Munroe (41).

Reprinted from *Psychol. Bull.*, 1949, 46, 393-429, by permission of the American Psychological Association and the author.

1. The writer wishes to express appreciation to Frederick Mosteller and to N. L. Gage, who read this manuscript and contributed many suggestions for its improvement.

In 1945, she considered the objectivity of previous Rorschach research. She distinguished between the goals attainable by clinical intuitive interpretation and the goals to be reached by more quantitative procedures. She traced the trend in Rorschach literature, noting the gradual decrease in studies based solely on impressionistic treatment of data or on mere counting of scores, and the introduction of significance tests, standard deviations, and other signs of adequate effort to test generalizations statistically. She also pointed out some errors in statistical thinking that lead to faulty conclusions about the Rorschach test. Munroe takes the position, and the writer fully concurs, that statistical research on the Rorschach test is not only justifiable, but indispensable. The flexibility of clinical thinking creates excellent hypotheses, but these hypotheses can be established as true only by controlled studies. Among the propositions suggested by clinical work, some are certainly untrue, due to faulty observation, inadequate sampling, and errors of thinking. Statistical controls are essential to verify theories of test interpretation, and to validate proposed applications of the test. Even though the clinician studying one person makes no use of statistics, he employs generalizations about the test which must rest on scientifically gathered evidence. Munroe demonstrated that the Rorschach test lends itself to objective studies; the writer reviews the same material more technically to evaluate the soundness of the statistical procedures on which the conclusions are based.

### *Clinical Treatments of Data*

While this paper deals principally with statistical methods applied to *raw* Rorschach data, we shall consider briefly the statistical procedures used when clinically interpreted case records are used in a study. The Rorschach record is usually interpreted qualitatively and in a highly complex manner when the test is given in the clinic, and many studies have been based on these interpreted records. In only a few studies of this type do statistical problems arise.

*Dichotomized Rorschach ratings.* In one type of study, the interpreter of the records makes a final summary judgment, dividing the records into such groups as "adjusted-maladjusted" or "promising-unpromising," etc. This method is most used for validation studies, where the Rorschach judgment is compared with a criterion of performance or with a judgment from some other test. Simple statistical tests suffice to test the degree of relationship. If the criterion is expressed in two categories (as when the criterion indicates success or failure for each case), chi-square is simple and appropriate. This



is exemplified in a study of success of Canadian Army officers (51), where a prediction from the Rorschach is compared with a later rating of success and failure. If the criterion is a set of scores on a continuous scale, bi-serial  $r$  is usually an adequate procedure. In bi-serial  $r$ , one assumes that the dichotomy represents a continuous trait which is normally distributed. This assumption is generally acceptable for personality traits and for ratings of success.

*Rorschach ratings on continuous scale.* In some studies, the Rorschach interpretation is reported in the form of a rating along a scale, rather than as a dichotomy. When the criterion is dichotomous, bi-serial  $r$  is appropriate. (E.g., a prediction of probable pilot success is so correlated with elimination-graduation from training, 21, p. 632.) For a continuous criterion, like grade-average, product-moment  $r$  is conventionally used.

These methods are not entirely satisfactory, because of a limitation of rating scales. If units on the rating scale are not psychologically equal, the correlation may not indicate the full size of the relationship. If ratings are careful, one can assume that men rated "Good" are superior to men rated "Fair," and that men rated "Excellent" are superior to both of these. But it may be unwise to assume that the jump from "Good" to "Excellent" is equal to the jump from "Fair" to "Good," as one automatically does in correlating. One solution to this difficulty is to assume that the trait rated is normally distributed in the men studied. Then we can condense the five-point scale into a dichotomy, which is the case discussed in the preceding paragraph. Alternatively, one may convert the ratings into scaled values which will yield a normal distribution (34). Bi-serial  $r$  is then appropriate, if the criterion is dichotomous. Similar reasoning applies to the correlation of a rating with a continuous criterion; one will obtain the most meaningful results by dichotomizing the rating and using bi-serial  $r$ , or by normalizing before using product-moment  $r$ . These suggestions are summarized in Table I.

Munroe (42), comparing a Rorschach adjustment rating with success in academic work, where both variables were reported on a four-category scale, used a coefficient of contingency. Where the correlation surface is nearly normal, this coefficient with proper corrections should give approximately the same result as the product-moment  $r$  for normalized data, corrected for broad categories. Yates (70) has recently offered an alternative method of adapting the contingency method to take advantage of trends in the relationship between variables expressed as ordered categories.

*Matching methods.* Another favorite technique for evaluating Rorschach results is blind matching, which permits a study of each case "as a whole." When a set of Rorschach records (interpreted or not) and another set of data regarding the same individuals are available, one may request

TABLE I  
Preferred Methods of Comparing Rorschach Interpretations with  
Criteria of Various Types

*Judgment Made From Rorschach*

Criterion	Dichotomy	Continuous Scale, Unequal Units
Dichotomy	$\chi^2$	$\chi^2$ after dichotomizing rating; $r_{bis}$ after normalizing rating*
Continuous scale, unequal units	$\chi^2$ after dichotomizing criterion; $r_{bis}$ after normalizing criterion*	$\chi^2$ after dichotomizing both variables; $r_{bis}$ after normalizing one, dichotomizing the other; product-moment $r$ after normalizing both
Continuous scale, equal units	$r_{bis}$ *	$r_{bis}$ after dichotomizing rating; product-moment $r$ after normalizing rating

\* Point bi-serial must be used if the two parts of the dichotomy cannot reasonably be considered subdivisions of a continuous scale.

judges to match the two sets in pairs. The success of matching is evaluated by a formula developed by Vernon (66). An example of its use is a study by Troup (62), in which judges tried to match two Rorschach records for each person. One hundred fourteen matches were correct out of a possible 120, judges considering five pairs at a time. By the Vernon formula, this corresponds to a contingency coefficient of .88. A coefficient of .40 was obtained when judges attempted to match the record of each case with that of his identical twin. Another excellent illustration of the method is provided by J. I. Krugman (31), who used it to establish that different evaluations of the same Rorschach protocol could be matched, and that the interpretations could be matched to the raw records and to criteria based on a case-study.

The limitations of this method are not statistical; they lie more in the human limitations of judges. A portrait based on the Rorschach may be nearly right, yet be mismatched because of minor false elements. Matching, on the other hand, might be excellent, even perfect; the study would still not guarantee that each element in each portrait was correct, especially if the subjects were quite different from each other. In fact, the portrait might be seriously wrong in some respects, without preventing matching.

A complex modification of the blind-matching method has been proposed and tried by Cronbach (9). Judges are asked to decide whether each statement on a list fits or does not fit a case described in a criterion sketch.



Since only about one-third of the statements in the list were actually made about the given case, one can test by chi-square whether the matching is better than chance. The method yields many interesting types of information: (a) an all-over estimate of the validity of predictions with relation to the criterion, (b) a separate estimate of the validity of the description for each case or for subgroups, and (c) an estimate of the validity of statements dealing with any one aspect of personality (e.g., social relations).

### *Errors in Statistical Studies*

The majority of statistical studies with the Rorschach test have treated Rorschach scores directly, with clinical judgment eliminated. This is an important type of investigation, which presents numerous problems. Before considering general questions of procedure, however, it is necessary to deal with several errors and unsound practices found in the literature reviewed. These miscellaneous errors must be pointed out lest they be copied by later investigators, and to suggest that the studies in which the errors occurred need to be re-evaluated.

*Significance tests for small samples.* The critical ratio is not entirely satisfactory when applied to small samples. When there are fewer than 30 cases per group, the  $t$  test is preferable. This would apply, for example, in Goldfarb's (19) comparison of obsessionals with supposedly normal adolescents. His significance ratios are a bit too high, since he used the formula  $\text{diff.}/\sigma_{\text{diff.}}$  with groups of twenty cases. (It may be noted also that Goldfarb's study does not permit sound generalizations about obsessionals as compared to other adolescents. The obsessionals had a mean IQ of 120 compared to 97 for the normals, so that differences between the groups may be due to intelligence rather than obsessional trends.)

Chi-square is generally useful for small samples, but it is important to apply corrections when the number of cases is below fifty. This is especially important when the expected frequency in any cell of a  $2 \times 2$  table is five or lower, under the null hypothesis. Many Rorschach studies fail to recognize the need for corrections, Kaback's (29, pp. 24, 38-39) being a striking example. She compares the distribution of such a score as  $M$  in each of two groups. To do so, she makes the distribution in a great number of intervals, with only a few cases per interval, and tests the similarity of the distributions by chi-square. In such a case, with many small cell frequencies, no significant result could be expected. Nor is it useful to inquire, as her procedure does, whether the precise distribution of  $M$  scores is the same for the two groups (in her case, pharmacists and accountants). Her major question was

whether one group used  $M$  more than the other, and this could be answered by dichotomizing the distribution and then applying chi-square, with proper correction. In applying chi-square to the  $2 \times 2$  tables, one should as a standard practice apply Yates's correction (56, p. 169). The importance of this correction will be demonstrated in Table IV. Where groups are dichotomized, it is best to make cuts toward the center, so that marginal totals will remain reasonably large. Special problems in the application of chi-square to successive tests of the same hypothesis, and to problems of goodness of fit, are discussed by Cochran (6).

*Tests for significance of difference in proportions.* Throughout the Rorschach literature, the formula for the significance of differences between proportions is misused. The resulting inaccuracy is slight in most problems, fortunately. This error is common in other work, and even some statistics books appear to endorse the faulty procedure. The usual formula,

$$\sigma_{p_1-p_2} = \sqrt{\frac{pq}{N_1} + \frac{pq}{N_2}}$$

may not be entered with  $p_1$  and  $p_2$ , the proportions obtained in the two samples. Instead, one should substitute  $p_0$  for  $p$ , where

$$p_0 = \frac{N_1 p_1 + N_2 p_2}{N_1 + N_2}$$

A significance test inquires whether  $p_1$  and  $p_2$  might arise by chance in sampling from a homogenous population in which the true proportion is  $p_0$  (see 35, pp. 126-129). Employing  $p_1$  and  $p_2$ , instead of entering  $p_0$  in both terms, almost always increases the critical ratio over what it should be. Because no correct model is found in the Rorschach literature, the following example is given using Hertz' data (25).

Five boys out of forty-one, and 0 girls out of thirty-five gave zero color responses.

$$p_0 = \frac{5 + 0}{76} = .066$$

$$s. d._{diff.} = \sqrt{\frac{.066 \times .934}{51} + \frac{.066 \times .934}{35}} = .057$$

$$p_1 = .122; p_2 = .00; \frac{\text{diff.}}{s. d._{diff.}} = \frac{.122}{.057} = 2.14 (P = .032)$$

This compares to the critical ratio of 2.41 ( $P = .016$ ) computed by the formula Hertz and other workers have inadvisedly used.



The above computation is equivalent to the determination of significance by chi-square, and yields an identical result. But in this instance the expected frequencies are so low that the correction for continuity becomes important. Applying Yates's correction, we find that  $P$  becomes .10, and the reported difference is not significant.

Several studies use the formula for proportions in independent samples when the formula for paired samples should be used. Thus Hertz (25), to compare the twelve-year-old and fifteen-year-old records of the same cases, should use a formula for correlated samples as given by Peatman (44, p. 407) or by McNemar (37; see also 13, 59). The correct formula would have yielded significant differences where Hertz found none. Other studies employing matched samples, where the significance of differences was underestimated by a formula for independent groups, are those of Hertzman and Margulies (27), Meltzer (39), M. Krugman (32), Richardson (48), and Goldfarb (20). In studies where the subjects were children varying widely in age, the proper formula would probably have yielded quite different results.

A study by Brown (4) committed this error and one even more serious. He compared records of twenty-two subjects without morphine and then with morphine. He found that fourteen increased in  $R$  and 7 decreased. He then treated these as independent proportions of the twenty-two subjects, computing the critical ratio for the difference 64% minus 32%. These are not proportions in independent samples, and Brown's statistical tests are meaningless. No manipulation of the increase-decrease frequencies is as satisfactory for this problem as the formula given by McNemar. Brown could properly have set a cutting score (e.g., 20R) and compared the percentage exceeding this level with and without morphine.

Siegel's procedure (55), in which the "percentage incidence" of a factor in one group is divided by the incidence in the second group, will be likely to produce misleading results.

An alternative formula for the significance of differences in matched groups is used by Gann (18). In applying the formula, however, a serious error was made. The formula given by Englehart which Gann adopted is

$$\sigma_{diff.}^2 = (\sigma_{M_1}^2 - \sigma_{M_2}^2)(1 - r_{if}^2)$$

$r_{if}$  is the correlation of the matching variables with the variable in which a difference is being tested. This formula may be extended to differences in proportions, although the estimated population value ( $p_0$ ) for the proportion should be substituted for  $M_1$  and  $M_2$ , as explained above. Gann's major error was to use a value of .9741 for  $r_{if}$  in all her calculations. From the context, this seems to be a multiple correlation of all matching variables with *all* dependent variables. The proper procedure, for any single significance

test such as the proportion of cases emphasizing *W*, would be to correlate the matching variables with *W*-tendency alone. This correlation would almost certainly be close to zero. By the procedure Gann used, the critical ratios are very much larger than they should be. In one comparison where Gann reported a CR of 6.02 the writer has established that the true CR cannot be greater than 2.23, and is almost certainly less.

*Comparisons of total number of responses.* It is thoroughly unsound to compare the total number of responses of a given type in two samples. Swift (58) tested thirty-seven boys and forty-five girls. The boys gave a total of 248 *F* responses; all girls combined gave 246. Swift used chi-square, demonstrating that these 494 responses were divided in a way which departs significantly from the theoretical ratio 37:45. But this assumes 494 independent events in her sample whereas she really had eighty-two. The *F* responses are not independent, since some were made by the same person. She might properly have used the *t*-test, applied to the means of the groups. The only correct way to use chi-square on her problem is to compare the number of cases exceeding a certain *F* score (cases, not responses, being the basis of sampling). A similar error has been made by Hertzman (26), Rickers (49, p. 231), and Werner (68).

Richardson (48) followed a different erroneous procedure. In her Table 9, she determined what proportion of all responses in each of her groups were *W* responses, and tested the difference in proportions for significance using the number of subjects in the denominator of the significance formula. The "proportion" she was studying is actually the ratio  $\text{Mean } W / \text{Mean } R$ , and the standard deviation of this is not correctly given by the formula  $\sqrt{pq/N}$ . If she must test the *W/R* ratio, in spite of the difficulties to be considered later, it is necessary to determine the ratio for each person separately and test differences between the groups in one of the conventional ways (e.g., chi-square, *t*-test, etc.).

*Inflation of probabilities.* Rorschach studies are peculiarly prone to an error which can arise in any statistical work. If a particular critical ratio or chi-square or *t*-test corresponds to a *P* of .05, we conventionally interpret that as statistically significant because "such a value would arise by chance only once in twenty times." While this usually refers to once-in-twenty samples, it may also be thought of as "once in twenty significance tests," if the several tests are independent. In some Rorschach studies, a vast number of significance tests are computed. Thus Hertz in one study reported the astonishing total of 800 significance tests (25). Many of these comparisons reach the 1 per cent level or the 5 per cent level, but even these are not all statistically significant. Quite a few of these differences did arise by chance, and unfortunately we cannot estimate how many because the tests were not



experimentally independent. The proper procedure, in such a case, is to recognize that an inflation of  $P$  values has taken place. The analogy to monetary inflation is a fair one: The increase in the number of significance tests in circulation causes each  $P$  to have less worth than it would normally. We may accordingly raise our "price" arbitrarily, and insist that  $P$  reach a higher level than .05 before we label it "significant," and a higher level than .01 before we label it "very significant." Of the differences reported in the Rorschach literature as "significant at the 5 % level," probably the majority are due to chance.

There are several ways in which significance levels may be inflated so that they become falsely encouraging. One is the common procedure of testing differences on a great many Rorschach scores. This is of course sound practice, but one must then take the total number of significance tests into account in evaluating  $P$ . The inflation is more subtle when the investigator rejects a large number of hypotheses by inspection without computing significance tests, and reports only a few significance tests. Thus Piotrowski and others (46) compared superior and inferior mechanical workers on "all the components used in conventional scoring as well as many others." They finally invented four composite scoring signs on which differences between the two samples were large enough to encourage a significance test. Suppose, for simplicity, that those four tests had yielded  $P$ 's of .02. The significance of those  $P$ 's must be minimized in view of the fact that four such differences were found in several hundred implied comparisons which were not actually computed, and two per hundred is chance expectation.

A comparable inflation arises when an investigator slices a distribution in order to take advantage of chance fluctuations and find some "hole" where a test will yield a low  $P$ . Hertz applied the formula for significance of the difference in proportions, to compare two groups on  $M\%$  (Table II). She introduced a spurious element by slicing the  $M\%$  distribution in so many places, and making so many significance tests. If a distribution is dichotomized in many ways, the chances of a "significant" difference rise greatly. Here only one test yielded a  $P$  of .05, out of nine attempted. The interpretation "It may be said with certainty, that more girls than boys at 15 years give over 11 %  $M$ " (25, p. 180) is unjustified. In another sample this fluctuation would not occur. It is not necessary to test explicitly all possible dichotomies for this type of error to arise. If the investigator examines his distribution and makes his cut at the place where the difference is greatest, he has by implication examined and discarded all other possible hypotheses. One of the several studies where this occurs is that of Margulies, discussed later.

TABLE II

Significance Data Reported by Hertz for Differences in *M%* Between  
Fifteen-Year-Old Boys and Girls (25)

Difference Tested	Critical Ratio	P
Difference in means	1.47	.15
Difference in medians	2.32	.05
Difference in proportions		
in interval 0-1	.81	—
in interval 0-3	.81	—
in interval 0-5	1.83	.10
in interval 0-7	1.68	.10
in interval 0-9	.90	—
in interval 0-11	2.34	.05
in interval 0-13	1.24	—
in interval 0-15	1.81	.10
in interval 0-17	1.23	—

Multiple correlation procedures give rise to a similar error. Suppose ten scores are tried as predictors. These scores might be combined in a prediction formula in an infinite number of ways. When an investigator computes correlations and works out the best possible predictive combination for his particular data, he implicitly discards all the other combinations. Even though his combination gives a substantial multiple *R* for the original sample, it is certain to give a lower correlation in a new sample where the formula can no longer take advantage of chance fluctuations. The common practice of comparing two groups on a large number of signs and developing a checklist score in which the person is allowed one point for every sign on which the two groups differ, is open to the same objection. In a new sample many of these signs will no longer discriminate.<sup>2</sup> When a significance test is applied to a difference in check list scores or to a multiple correlation in the sample on which the combining formula was derived, the significance test has only negative meaning. If, even after taking advantage of chance differences, one's formula cannot discriminate, it is indeed worthless. But if the result gives a *P* better than .05, the formula may still be of no value. Rorschach studies which have reported "significant" differences based on an empirical formula without confirming them on fresh samples are those of Montalto (40), Harris and Christiansen (23), Hertzman, Orlansky, and Seitz (28), and Ross and Ross (52). Thompson (60) reports spurious *r*'s but does not claim significance for them. Buhler and Lefever (5, Tables X,

2. Harris (24) claims that in his experience the Rorschach behaves differently from other tests, and that signs found to differentiate in one sample are usually confirmed in other samples. This appears improbable on logical grounds, and no evidence in the literature supports such a statement.



XX) mix new cases with the sample used in deriving scoring weights, and therefore fail to provide an adequate test of significance. Significance tests on fresh samples have been properly made by Guilford (21), Gustav (22), Margulies (38), Ross (50), and Kurtz (33). The latter gives a particularly clear discussion of the issue involved. In most studies, correlations nearly vanish when a Rorschach prediction formula is tried on a new sample.

Still another method of inflating probabilities is to recombine groups of subjects in a way to maximize differences. If one has several types of patients, all of whom earn different mean *M* scores, these groups may be recombined in many ways, and in one of the possible regroupings a pseudo-significant difference may be found. Rapaport and his coworkers (47) have carried inflation to bizarre levels. Not only did they consider scores in great profusion and in numerous combinations, they recombined their subjects so that the number of implicit significance tests in their volume is incalculable. They began with subjects in twenty-two subgroups. Significance tests were then made, on any score, between any pair of subgroups or combinations of them which seemed promising *after* inspection of the data. There were 231 possible pairs of subgroups, and an endless variety of combinations. Thus at times Unclassified Schizophrenics Acute were lumped with one, two, or more of the following: Paranoid Schiz. Acute, Simple Schiz., Uncl. Schiz. Chronic, Par. Schiz. Chr., Uncl. Schiz. Deteriorated; or with all the schizophrenics and preschizophrenics; or with Paranoid Condition, Coarctated Preschiz., Overrideational Preschiz., and Obsessive-Compulsive Neurosis. Such willingness to test any hypothesis whatever leaves these workers open to the charge of having regrouped their cases to augment differences. They have undoubtedly reported differences which were created by artificial combinations of chance variations between groups. Every time cases are recombined for a significance test, one must recognize that a large number of implied significance tests were also made, since many other recombinations were rejected without actual computation.

Rorschach studies, because of the great number of scores and the large number of subgroups of subjects involved, are more prone to inflation than other research. The suggestions to be made for sound practice are these:

1. Compare the number of significant differences to the total number of comparisons in the study, both those computed and those rejected by implication.
2. Raise the *P* value required for significance as the number of comparisons increases.
3. Never accept an empirical composite score or regression formula until its discriminating power has been verified on a new sample.
4. In general, do not trust significance unless the hypothesis tested was set up independent of the fluctuations of a particular sample.

These suggestions require that the investigator have clearly in mind the number of comparisons considered. Comparisons are of three types: those rejected as improbable before the data are looked at, i.e., before the study is begun; those not computed because a cursory inspection showed no apparent difference; and those computed. Sometimes the investigator begins with, say, five groups of subjects and ten scores, and frankly wants to unearth all possible differences between types of subjects. Then there are ten ways the groups may be paired against each other, and since each pair may be compared on each score, there are a total of one hundred comparisons in the study. If, on the other hand, the investigator sets out to check only certain relationships—"Schizophrenics differ from neurotics in  $F + \%$ ," "Manics differ from all other groups combined in  $FC : CF + C$ "—those limited hypotheses may be laid down in advance of the study, and only those comparisons are counted as implied significance tests. To avoid confusion, it is also well for the investigator to specify his cutting point, if a variable is to be dichotomized, before examining the differences between groups. This may be set by an arbitrary rule, for instance that each distribution is to be divided as near to its median as possible, or by an *a priori* decision to divide at some point such as  $2M$ . In essence, the investigator must ask himself before he gathers his data, "How many comparisons do I intend to look at, and charge myself for?" A  $P$  of .01 may be called significant if it is one of three comparisons charged for, but not if the investigator has looked at 300 comparisons in order to salvage this one impressive value.

### *Methods of Comparing Groups on Rorschach Scores*

#### NECESSITY FOR CHOOSING BETWEEN STATISTICAL PROCEDURES

Because Rorschach scores are numbers which can be added, averaged, distributed, etc., most investigators have used conventional mental-test statistics without question. The most common need for statistics is to compare the test scores of groups and determine the significance of differences. The prominent methods encountered in Rorschach literature are as follows: significance of difference between means (critical ratio or  $t$ -test); analysis of variance; bi-serial  $r$ ; significance of difference in proportions exceeding a particular score, or chi-square; and significance of difference between medians.

Apart from such errors as those listed in the preceding section, there is



no reason for considering any of the procedures under discussion as mathematically incorrect. If a significant difference is revealed by any proper significance test, the null hypothesis must be rejected. Nevertheless, the investigator may not choose one of the techniques at random. *Different methods of analyzing the data will lead to different conclusions.* In particular, some procedures lead to a finding of no significant difference even though a true difference could be identified by another attack.

Let us illustrate first with some of Kaback's data (29). She administered the group Rorschach to men in certain occupations, and, *inter alia*, compared her groups on the number of popular responses. The means for accountants is 7.0; for accounting students, 7.3. By the *t*-test, the difference between means is not significant ( $P$  ca. .40). (Point bi-serial  $r$  applied to the same data gives the same significance level. Point bi-serial  $r$  and  $t$  are interchangeable procedures, and there is no merit in testing the hypothesis in both ways.) But if she had chosen the chi-square test, quite proper for her data, Kaback would have found a significant difference between the groups. Chi-square would be applied to compare the proportion of cases in each group having five or more popular responses. From her Table IV, this proportion is 60/75, accountants; 72/75, accounting students. The difference between accountants and accounting students is significant ( $P < .01$ ). In this and other instances, Kaback disregarded a difference when the null hypothesis could be confidently rejected.

Further illustrative data are taken from Hertz' comparison of Rorschach scores of boys and girls. She tested each possible difference by several statistical devices, yielding results such as those for  $M\%$  reproduced in Table II. By any of nine methods, she is informed that the two sex groups differ no more than might two chance samples. By the other computations, she is informed that the difference is significant at the 5% level. If different significance tests disagree, what one concludes depends nearly as much on what procedure one adopts as on the data themselves.

Hertz compared her boys and girls in forty-six instances. Each time, she tested the significance of differences between means and between medians. Four times the means differed significantly; five times, the medians differed significantly. But in only one out of forty-six comparisons was the difference significant by both methods. It is greatly to Hertz' credit that she saw the applicability of more than one significance test. But conclusions of research will be hopelessly confused and contradictory, unless we can find a basis for choosing between the procedures when one says "'Tis significant" and the other says "'Taint."

The choice between comparison of means and medians or between the *t*-test and chi-square cannot be left to the inclination of the experimenter;

the whole point of statistical method is to make an analysis freed from subjective judgment. The reason different methods yield different results is that they make different assumptions or try to disclose different aspects of the data. It is therefore important to recognize the ways in which the techniques differ. Differences which are of little concern in connection with most studies have peculiar importance in Rorschach work. The difficulties which make choice of procedures an important problem arise from three causes: the skewness of Rorschach scores, the complications introduced by ratio scores, and the dependence of Rorschach scores on the total number of responses.

#### CHOICE OF TECHNIQUES IN VIEW OF THE INEQUALITY OF UNITS IN RORSCHACH SCALES

Many of the significant Rorschach scores give sharply skewed distributions for most populations. This fact is reported repeatedly (2, 25, 47). Skewness is usually found where many subjects earn 0, 1 or 2 points (i.e.,  $M$ ,  $FM$ ,  $m$ , the shading scores,  $CF$ , and  $C$ ), and in the location scores  $W$ ,  $D$ ,  $Dd$ , and  $S$ . Skewness itself is no bar to conventional significance tests. But in skew distributions the mean and median are not the same. Two distributions may have a significant difference in medians, and not in means (or vice versa) if either is skewed.<sup>3</sup> Furthermore, it is doubtful if a satisfactory estimate of  $s.d._{mdn}$  can be obtained for a skewed distribution.

*Disadvantages of the mean and related procedures.* In any statistical computation based on addition of scores (mean  $s.d.$ ,  $t$ , analysis of variance), numerical distances between scores at different parts of the scale are treated as equal. Thus, since the average of 3  $W$  and 7  $W$  is the same as that of 1  $W$  and 9  $W$ , these computations assume that a shift 3  $W$  to 1  $W$  is equivalent to, or counterbalances, a shift 7  $W$  to 9  $W$ . There is no way of demonstrating equality of units unless one has some knowledge of the true distribution of the trait in question, or a definition of equality in terms of the characteristics of the property being measured. This problem is present in virtually all psychological tools, but other tests yield normal distributions which are assumed to represent the true spread of ability. On the other hand, *Rorschach interpretation based on clinical experience constantly denies the equality of units for Rorschach scores.* The average  $W$  score is near 6, and scores from 1 to 10 are usually considered to be within the normal range. No matter how extremely a person is lacking in  $W$  tendency, his score cannot go below zero. For one who overemphasizes  $W$ , the score may go up to 20, 30, or more. A  $W$  score only six points below the mean may be consid-

3. This argument is presented by Richardson (48). In attempting to study differences in medians, Richardson unfortunately uses an incorrect method of determining  $s.d._{mdn}$ .



ered clinically to be as extreme in that direction as a score fifteen points from the mean in the other direction. Munroe (42) has prepared a checklist which shows how units of certain Rorschach scores would have to be grouped in order to represent a regularly progressing scale of maladjustment. Her groupings based on clinical experience are of approximately this nature:

$W$  (or  $W\%$ ): 0 (or 1 poor)  $W$  response; 1-14%; 15-60%; 61-100%.

$Dd\%$ : 0-9%; 10-24%; 25-49%; 50-100%.

$m$ : 0-1; 2-3; 4-5; 6 or more.

If these units represent increasing degrees of maladjustment, the raw Rorschach scores do not form a scale of psychologically equal units. It is advisable to accept the clinical judgment on this point, especially in the absence of evidence for the assumption of equal units.

*Use of median and chi-square.* Unlike procedures involving the addition of scores, procedures based on counting of frequencies make no assumption about scale units. In fact, they give the same results no matter how the scale units are stretched or regrouped. The median, or the number of cases falling beyond some critical point (e.g. 10  $W$ ), depends only on the order of scores. This appears to justify the recommendation that counting procedures such as the median be given preference over additive procedures such as the mean in dealing with skew Rorschach distributions. To test the significance of a difference between two groups, the best procedure is to make a cut at some suitable score, and compare the number of cases in each group falling beyond the cut, using chi-square. This procedure is used by Rapaport (47) and Abel (1). The test of significance of differences between proportions yields the same result (see above). One virtue of cutting scores is that we may test for differences between groups both in the "high" and "low" directions. This is important, since either very high  $F\%$  or very low  $F\%$ , for example, may have diagnostic significance. In the usual analysis based on means, deviations of the two types cancel.

In contrast to the chi-square method, many tests of significance involve computation of the standard deviation. These include the critical ratio of a difference between means or medians, analysis of variance, and the  $t$ -test. In these procedures, great weight is placed on extreme deviations from the mean. If mean  $W$  is 6, a case having 25  $W$  increases  $\Sigma d^2$  (which enters the computation of the  $s.d.$ ) by about 361 points; a case having 15  $W$  increases  $\Sigma d^2$  by about 81 points; and 0  $W$ , by only 36 points. In skewed Rorschach distributions, the few cases with many responses in a category have a preponderant weight in determining  $\sigma$  and the significance of the difference.

Whether weighting extreme cases heavily is acceptable depends on whether one considers the difference between 15 *W* and 25 *W* to be psychologically large and deserving of more emphasis than, say, the difference from 0 *W* to 5 *W*. Chi-square weights equally all scores below (or above) the cutting point.

*Normalizing distributions.* One method used to obtain more equal units is to assume that the trait underlying the score is distributed normally in the population. Raw scores are converted to *T*-scores which are normally distributed (35, 67). (This procedure must be distinguished from another conversion, also called a *T*-score, used by Schmidt (54). Scores of the type Schmidt used are not normally distributed.) The effect of normalizing is to stretch the scale of scores as if it were made of rubber. Extreme scores below the median are weighted symmetrically to extreme scores above the median. Thus, in the conversion table prepared by Rieger and used by the writer (10), the median ( $61\frac{1}{2}$  *W*) is placed at 100, and a score of 0 *W* is converted to 66, while 28 *W* becomes 134. This in effect compresses the high end of the *W* scale and expands the low end. This conversion does not alter any conclusion or significance test obtained by dichotomizing raw scores and applying chi-square. But the conversion alters markedly any conclusion based on variance or on comparison of means.

There is obviously much merit in using a procedure which leads to a single invariant result, independent of the assumption of the investigator about the equivalence of scores. Even if scores are normalized it is advised that the median be used to indicate central tendency, and chi-square to test significance. If, for some experimental design, the data must be treated by analysis of variance, the writer believes normalized scores will give results nearer to psychological reality than raw scores, but this judgment is entirely subjective.

*Comparison of mean rank.* Attention should be drawn to a new technique invented by Festinger (14) which is peculiarly suitable to the problem under discussion. This method assumes nothing about equality of units or normality of distributions, being based solely on the rank-order of individuals. To test whether two groups differ significantly in a score, one pools the two samples and determines the rank of each man in the combined group. The mean rank for each group is computed and the significance of the difference is evaluated by Festinger's tables. The method has not yet been employed in Rorschach research.

The Festinger method and chi-square are not interchangeable. Which should be used depends on the logic of a particular study. Chi-square answers such a question as "Does Group A contain more *deviates* than Group B in the score being studied?" The Festinger method gives weight to differences all along the scale, and therefore asks whether the two groups



differ, all scores being considered. In one study, absence of *M* is quite important but differences in the middle of the range have no practical importance. In another study, differences all along the scale are worth equal attention.

The Festinger method appears to have the advantage of greater stability for small samples. Chi-square is much easier to use in samples of 30 or more per group. The Festinger method is not useful when there are numerous ties in score. Further experience with the new method may disclose other important distinctions.

#### SIGNIFICANCE TESTS COMPARED WITH ESTIMATES OF RELATIONSHIP

Some investigators have perhaps not conveyed the full meaning of their findings to the reader because of a failure to distinguish between tests of the null hypothesis, and estimates of the probable degree of relationship between two variables. The former type of result is a function of the number of cases, whereas the latter is not, save that it becomes more trustworthy as more cases are included. When an investigator applies chi-square, the *t*-test, or the like, he determines whether his observations force him to conclude that there is a relationship between the variables compared. But if the degree of relationship is moderately low, and the number of cases small, the null hypothesis is customarily accepted even though a true relationship exists. It is proper scientific procedure to be cautious, to reject the hypothesis of relationship when the null hypothesis is adequate to account for the data. But in Rorschach studies, where sample size has often been extremely restricted, nonsignificant findings may have been reported in a way which discouraged investigators from pursuing the matter with more cases.

The study of McCandless (36) is a case in point. McCandless compared Rorschach scores with achievement in officer candidate school. In each instance save one, the *t*-test showed *P* greater than .05 that the difference would arise in chance sampling. But the samples compared contained only thirteen men per group. Under these circumstances, it would take a sharply discriminating score to yield a significant difference. If the sample size was raised to about fifty per group, and the differences between groups remained the same, twelve more of McCandless' thirty significance tests would be significant at the five percent, or even the one percent, level. When more cases are added, the differences will certainly change and most of them will be reduced in size. In fact, the writer believes, on the basis of other experience with statistical comparisons of the Rorschach with grades, that McCandless' negative findings are probably close to the results which would be

found with a larger sample. But the point is that McCandless, and other investigators using small  $N$ 's, have submitted the Rorschach to an extremely, perhaps unfairly, rigorous test. One way to compensate for the necessary rigor of proper significance tests is to also report the degree of relationship. A chi-square test may be supplemented by a contingency coefficient or a tetrachoric  $r$ . A  $t$ -test may be supplemented by a bi-serial  $r$ , or point bi-serial (not to determine significance, as Kaback used it, but to express the magnitude of the relationship). Sometimes reporting the means of the groups and their standard deviations, to indicate the degree of overlapping, is an adequate way to demonstrate whether the relationship looks promising enough to warrant further investigation.

To restate the problem: the investigator always implies two things in a comparison of groups: (a) that he considers the null hypothesis is definitely disproven by his data, or else that the null hypothesis is one way to account for the data, and (b) in case the null hypothesis still remains tenable, that he does or does not judge further investigation of the question to be warranted. He can never prove that there is no relationship. So, if his data report a non-significant difference, he must judge whether the difference is "promising" enough to warrant further studies. This judgment is not reducible to rules in the way the significance test is. Whether to recommend further work depends on the difficulty of the study, on the probable usefulness of the results if a low order of relationship were definitely established by further work, and in the investigator's general confidence that the postulated relationship is likely to be found.

#### METHODS OF PARTIALLING OUT DIFFERENCES IN R

The usual approach when comparing groups is to test the differences in one score after another, and then to generalize that the groups differ in the traits to which the scores allegedly correspond. The various scores, however, are not experimentally independent—a man's total record is obtained at once, and his productivity influences all his scores. If two groups differ in  $R$ , they may also differ in the same direction in  $W$  (whole responses),  $D$  (usual details), and  $Dd$  (unusual details).

Thus consider the Air Force data in Table III.

The first group has more responses than the second. From the means in  $W$  and  $D$ , it would appear that the first group has more  $W$  tendency than the second, but is equal in  $D$ . But when responsiveness is controlled by converting scores to percentages, the difference in  $W$  becomes small and the second group is shown to be stronger than the first in emphasis on  $D$ .



TABLE III

Rorschach Scores Compared to Success in Pilot Training (21, p. 632)

Rorschach Score	Mean of Successful Cadets	Mean of Unsuccessful Cadets	Bi-serial $r$
<i>R</i>	18.5	15.8	.14
<i>W</i>	9.2	7.3	.24
<i>D</i>	7.1	6.7	.03
<i>W</i> %	60.2	55.8	.08
<i>D</i> %	31.7	37.6	-.15

The most striking illustration of this difficulty is Goldfarb's comparison of obsessionals and normals. The obsessional group averages fifty-five *R*; the normals, fourteen. Under the circumstances, it is not at all informative to proceed to test *W*, *D*, and *Dd*; all differ significantly in the same direction. One learns nothing about differences between groups in mental approach, which is the purpose of considering these three scores. Most of Goldfarb's other comparisons also merely duplicate the information given by the test in *R*, that is, that the obsessionals are more productive. Although the discrepancy between the groups in *R* is unusually striking in Goldfarb's group, it is present to a lesser but significant degree in a great number of other studies, including those of Buhler and Lefever (5), Hertzman (26), Kaback (29), Margulies (38), and Schmidt (54).

A similar problem complicated Beck's comparison of schizophrenics and normals on *D*. The means were 19.0 and 19.9, respectively; the  $\sigma$ 's were 13.5 and 9.9. Beck comments as follows:

The small difference is accentuated in the very small Diff./S.D. diff. 0.34. There is, however, probably a spurious factor in this small difference. The ogives give us a hint; up to the eighty-second percentile, the curves run parallel, with that for controls where we should expect it, higher. Above this point, the schizophrenics' curve crosses over, and continues higher, and more scattering, as we should expect from the S.D. The spurious element lies undoubtedly in the fact that the schizophrenics' higher response total would necessarily increase the absolute quantity of *D*, since these form the largest proportion of responses in practically all records. Absolute quantity of details is then no indicator of the kind of personality we are dealing with. . . . The medians for *D* are 14.46, 17.2 (2, pp. 31-32).

When one makes several significance tests in which the difference in *R* reappears in various guises, one becomes involved in a maze of seemingly contradictory findings. And interpretation tempts one to violate the rule of parsimony, that an observed difference shall be interpreted by the fewest and simplest adequate hypotheses. To answer the question, how do ob-

sessionals and normals differ? it is simpler to speak of the former as more productive than to discuss three hypotheses, one for each approach factor. And one may certainly criticize Hertzman and Margulies (27) for interpreting differences in  $D$  and  $Dd$  between older and younger children as showing the former's greater "cognizance of the ordinary aspects of reality" and greater concern with facts. The older group gives twice as many  $R$ 's as the former, which is sufficient to account for the remaining differences.

One might argue that  $R$  is resultant rather than cause, and that the differences in  $W$ ,  $D$ ,  $Dd$ , etc., are basic. But the Air Force demonstration that  $R$  varies significantly from examiner to examiner (21) suggests strongly that responsiveness is a partly superficial factor which should be controlled.

Only two studies examine their data explicitly to determine if differences in other categories could be explained in terms of responsiveness alone. Werner (68) found a significant difference in  $dd\%$  between brain-injured and endogenous defectives. But the latter gave significantly more  $R$ 's. He therefore counted only the first three responses in each card, and arrived at new totals. With  $R$  thus held about constant, he found the  $dd$  difference still marked and could validly interpret his result as showing a difference in approach.

Freeman and others (17) found that groups who differed in glucose tolerance also differed significantly in  $R$ . After testing differences in  $M$  and  $\text{sum } C$  on the total sample, they discarded cases until the two subsamples were equated in  $R$ . Since differences between the groups in  $M$  and  $C$  were in the same direction even when  $R$  was held constant, they were able to conclude with greater confidence that glucose tolerance is related to  $M$  and  $C$ .

After differences in  $R$  are tested for significance, it is appropriate to ask what other hypotheses are required to account for differences in the groups. But these other hypotheses should be independent of  $R$ ; otherwise one merely repeats the former significance test and obscures the issue. The usual control method is to divide scores by  $R$ , testing differences in  $W\%$ ,  $D\%$ ,  $M\%$ ,  $A\%$ ,  $P\%$ , etc. Such ratios present serious statistical difficulties discussed in the next section. Moreover, these formulas fail to satisfy the demand for independence from  $R$ . There may be correlation between  $R$  and  $W\%$ , etc. (For a sample of 268 superior adults from a study by Audrey Rieger, the writer calculates these  $r$ 's:  $W\% \times R$ ,  $-.45$ ,  $M\% \times R$ ,  $.03$ ,  $F\% \times R$ ,  $.06$ . In the latter two cases, there is no functional relation of the percentage with  $R$ , but the distributions are heteroskedastic.  $\sigma_{w\%} = 3.30$  when  $R$  5-19 (74 cases) but 2.09 when  $R$  40-109 (82 cases). The corresponding sigmas for  $M\%$  are 3.85 and 3.35; for  $F\%$ , 3.23 and 2.29. Only  $M\%$  is really independent of  $R$ .)

One may control differences in  $R$  by other methods, provided many



cases are available. One procedure is to divide the samples into subgroups within which  $R$  is nearly uniform (e.g.,  $R$  20–29), and make significance tests for each such set. A method which requires somewhat fewer cases is to plot the variable against  $R$  for the total sample or a standard sample, and draw a line fitting the medians of the columns. This may be done freehand with no serious error. Then the proportion of the cases in each group falling above the line of medians may be compared by chi-square.

#### DIFFICULTIES IN TREATING RATIOS AND DIFFERENCES

More than any previous test in widespread use, the Rorschach test has employed "scores" which are arithmetic combinations of directly counted scores. One type is the ratio score, or the percentage in which the divisor is a variable score. Examples are  $W:M$ ,  $M:\text{sum } C$ ,  $W/R$  ( $W\%$ ), and  $F/R$  ( $F\%$ ). The other type of composite is the difference score, such as  $FC - (CF + C)$ . In clinical practice, scores of this type are used to draw attention to significant combinations of the original scores; the experienced interpreter thinks of several scores such as  $FC$ ,  $CF$ , and  $C$ , at once, placing little weight on the computed ratio or difference. When these scores are used statistically, however, there is no room for the flexible operation of intelligence; the ratios are treated as precise quantities.

It may be noted in passing that a few workers (e.g., 63) appear to assume that Mean  $a/\text{Mean } b$  is the same as Mean  $\frac{a}{b}$ . This is of course not true; the mean of the ratios and the ratio of the means may be quite unequal. One cannot, as Kaback did (29, pp. 33, 53, 55), assume that if the ratio of the means is greater for one group than another, the groups differ in the ratio scores themselves. The reader may convince himself by computing the mean ratio for each of the following sets of data in which Mean  $a/\text{Mean } b$  is constant:

$$\frac{0}{2}, \frac{2}{4}, \frac{4}{6}, \frac{6}{8}, \frac{8}{10}; \quad \frac{0}{6}, \frac{2}{8}, \frac{4}{10}, \frac{6}{12}, \frac{8}{14}; \quad \frac{0}{10}, \frac{2}{8}, \frac{4}{6}, \frac{6}{4}, \frac{8}{2}.$$

One difficulty with ratio scores is their unreliability. Consider a case with 5  $W$ , 1  $M$ . The ratio  $W:M$  is 5. But  $M$  is a fallible score. On a parallel test it might shift to 0 or to 2. If so, the ratio could drop to  $2\frac{1}{2}$ , or zoom to infinity; such a score is too unstable to deserve precise treatment. The unreliability of another ratio is illustrated in Thornton and Guilford's data (61). The reliabilities were, in one sample, .92 for  $M$ , .94 for  $C$ , but .81 for  $M/C$ . In a second sample, the values were .77, .65, and .31. If unreliable ratios are added, squared, and so on, one commits no logical error, but psy-

chologically significant differences become overshadowed by errors of measurement.

Ratios based on small denominators are in general unreliable (7).  $W\%$  is unreliable for a subject whose  $R$  is 12, but relatively reliable for a case whose  $R$  is 30. In the former case, addition of one  $W$  response raises  $W\%$  by 8; in the latter, by 3%. Errors of measurement always reduce the significance of differences by increasing the within-groups variance. A significant difference in  $W\%$  might be found for cases where  $R > 25$ . A difference of the same size might not be significant for cases where  $R < 25$  because of the unreliability of the ratio. If the significance test were based on all cases combined, the difference might be obscured by the unreliability of the ratios in the latter group. One possible procedure is to drop from the computations all cases where the denominator is low. (If there is a significant difference even including the unreliable scores, this need not be done.)

The issue of skewness must again be raised. In the  $M:sum C$  ratio, all cases with excess  $C$  fall between zero and 1. Those with excess  $M$  range from 1 to  $\infty$ . The latter cases swing the mean and sigma. Following the argument of a preceding section, it is injudicious to employ statistics based on the mean and standard deviation, as McCandless (36) did. By such procedures, different conclusions would often be reached if both  $M:sum C$  and  $sum C:M$  were tested. Procedures leading to a chi-square test are to be recommended, as illustrated in several studies (Rapaport, 47, pp. 251; Rickers, 49; etc.) Another solution, less generally suitable, is to convert ratio scores to logarithmic form to obtain a symmetrical distribution (61).

A hidden assumption in ratios and differences is that patterns of scores yielding equal ratios (or differences) are psychologically equal. Thus, in  $W\%$  the same ratio is yielded by 2  $W$  out of 10  $R$ , 8  $W$  out of 40, and 20  $W$  in 100  $R$ . One can always define and manipulate any arbitrary pattern of scores without justifying it psychologically, but better conclusions are reached if the assumption of equivalence is defensible. The regression of  $W$  on  $R$  is definitely curved. A person with 2  $W$  out of 10  $R$  is low in  $W$  tendency, since it is very easy to find two wholes in the cards. Only people with strong tendency and ability to perceive wholes can find 20  $W$  in the ten cards, regardless of  $R$ . As  $R$  rises above 40,  $W$  seems to rise very little; the additional responses come principally from  $D$  and  $Dd$ . The resulting decline in  $W\%$  reflects a drive to quantity, rather than a decreased interest in  $W$  (cf. 47, p. 156). Put another way: a strong drive to  $W$  can easily lead to 90 or 100%  $W$  when  $R < 15$ ; but such a ratio in a very productive person is unheard of. If the regression of  $a$  on  $b$  is linear and a close approximation to  $(a/b) = \text{some constant}$ , ratios may be used as a score with little hesitancy. Otherwise the ratio is a function of the denominator.



This factor is recognized by Munroe, who indicates repeatedly in her checklist that the significance of a particular ratio depends on  $R$ . Thus 30–40%  $M$  is rated + if  $R = 10$ , but 16–29% is rated + if  $R = 50$ . Numerically equal Rorschach ratios, then, are not psychologically equal. Rapaport reflects the same point in testing differences between groups in  $W/D$ . Instead of applying chi-square to the proportions having the ratio 1:2 or lower, he adjusted his standard.

In records where  $R$  is too low or too high, we took cognizance of the fact that it is difficult not to get a few  $W$ 's and difficult to get too many. Thus, in low  $R$  records the 1:2 norm shifted to a "nearly 1:1" while in high  $R$  records, the 1:2 norm shifted to a 1:3 ratio (47, p. 134).

This adjustment was evidently done on a somewhat subjective basis, and is therefore not the best procedure. It is unfortunate that most other workers have unquestionably assumed that a given score in  $W\%$ ,  $M\%$ , or  $FC - (CF + C)$  has the same meaning regardless of  $R$ .

At best, ratio- and difference-scores introduce difficulties due to unreliability and to assumptions of equivalence. There is a fairly adequate alternative which avoids statistical manipulation of ratios entirely. One need only list all significant patterns, and determine the frequency of cases having a given pattern. Thus  $M:sum C$  can be treated in these categories: coarctated ( $M$  and  $C$  2 or below); ambiequal,  $M$  or  $C < 2$ ,  $M$  and  $C$  differ by 2 or less; introversive,  $M$  exceeds  $C$  by 3 or more; extratensive,  $C$  exceeds  $M$  by 3 or more. Any other psychologically reasonable division of cases may be made, and significance of differences tested by chi-square, provided that the hypothesis is not chosen to take advantage of fluctuations in a particular sample. Even this method, however, does not escape the criticism that a given pattern of two scores, such as 3  $M$ , 3  $C$ , has different significance in records where  $R$  differs greatly. To cope with this limitation, the pattern tabulation procedure is suggested later.

A detailed consideration of certain work by Margulies is now appropriate, since it affords an illustration of many problems presented above. Her study of the  $W:M$  ratio employed a procedure almost like that just recommended, but with departures which are unsound. Margulies compared Rorschach records of adolescents having good and poor school records (38). Only her twenty-one successful boys and her thirty-two unsuccessful boys need be considered here. She was interested in comparing them on the  $W:M$  pattern, in view of Klopfer's belief that this ratio indicates efficient or inefficient use of capacity. She not only tested her data in several ways, but reported the data so that other calculations can be made. Table IV reproduces a part of her data, and shows the results of seven different procedures for determining the significance of the difference.

TABLE IV

Results Obtained When a Set of Data is Treated by a Variety of Procedures  
(Data from Margulies, 38, pp. 23, 26, 44)

<i>Distribution I</i>			<i>Distribution II</i>			<i>Distribution III</i>		
Number of <i>M</i>	Suc- cessful Boys	Unsuc- cessful Boys	<i>W/M</i> Ratio	Suc- cessful Boys	Unsuc- cessful Boys	Pattern of <i>W</i> and <i>M</i>	Suc- cessful Boys	Unsuc- cessful Boys
3 or more	5	5	<1	1	1	<i>W</i> <6, <i>M</i> 0-1	0	10
2	9	8	1.00	0	2	<i>W</i> <6, <i>M</i> >1	8	2
1	3	11	1.1-2.9	8	5	<i>W</i> >5, <i>M</i> 0-1	7	9
0	4	8	3.0-4.9	5	7	<i>W</i> 6-10, <i>M</i> 2	3	7
			>4.9	3	9	<i>W</i> 6-10, <i>M</i> >2	1	4
			$\infty(W/0)$	4	8	<i>W</i> >10, <i>M</i> >1	2	0

It should be noted first that Yates's correction is essential for tables with 1 d.f. and low frequencies; in each case where it is applicable, the correction lowers the significance value importantly. Second, attention may be turned to the use of chi-square to test differences between two distributions. Even if more cases were available, it would be unwise to apply chi-square to the distribution cell by cell (Procedures 2, 3), since this procedure ignores the regular trend from class-interval to class-interval. Instead, the distribution should be dichotomized. Therefore, procedure 5 is preferable to 2, and 6 is preferable to 3. It will be noted that these recommended procedures indicate higher significance than the tests in which the distributions are compared cell by cell.



TABLE IV (continued)

Type of Analysis	Procedure	Result	P	Results with Yates's Correction	
				$\chi^2$	P
Central tendency	1. Significance of difference in mean $M$	CR = .70*	.48	..	..
Cell-by-cell comparison	2. Chi-square applied to Distribution I (3 d.f.)	$\chi^2 = 3.78^{***}$	ca. .30	..	..
	3. Chi-square applied to Distribution II (5 d.f.)	$\chi^2 = 5.30^*$	ca. .40	..	..
	4. Chi-square applied to Distribution III (5 d.f.)	$\chi^2 = 17.73^*$	<.01	..	..
Dichotomy	5. Chi-square applied to number of cases with $M > 1$ (Dist. I)	$\chi^2 = 3.46^{**}$	.06	2.54**	.11
	6. Chi-square applied to number of cases with $W/M > 3$ (Dist. II)	$\chi^2 = 1.86^{**}$	.18	1.13**	.30
Frequency of selected patterns	7. Chi-square applied to frequency having $M > 1$ if $W > 6$ or $> 10$ ; having $M > 2$ if $6 < W < 10$ (Dist. III)	$\chi^2 = 6.58^{**}$	.01	5.13**	.03

\* Computed by Margulies.

\*\* Computed by the writer.

\*\*\* Computed by the writer. Margulies reports 3.64.

Margulies is one of the few writers to note the unsoundness of assuming that equal ratios are equal. She pointed out that 20  $W$ : 10  $M$  is not psychologically similar to 2  $W$ :1  $M$ , and she demonstrated that the regression of  $M$  on  $W$  is significantly curvilinear. She therefore was properly critical of procedures such as 3 and 6. She next turned to the scatter diagram of  $M$  and  $W$ , and found successful boys predominating in some regions, and unsuccessful boys in others. After grouping scores into regions as shown in Distribution III, she divided the surface into two areas, one area including cases where  $W$  is 0 to 5 and  $M$  is 2 or over, plus cases where  $W$  is 6 to 10 and  $M$  is 3 or over, plus cases where  $W$  is over 10 and  $M$  is 2 or over. In other words, instead of testing whether the groups are differentiated by a cut along

the straight line  $M = 2$  (Procedure 5), she made her cutting line an irregular one. This hypothesis, tested in Procedure 7, gave apparently quite significant results. The results are of little value, however, since the hypothesis was "cooked up" to fit the irregularities of these specific data. In the cells where  $W$  is 6 to 10, and  $M$  is 2, there happens to be a concentration of unsuccessful boys. But to draw the cutting line irregularly to sweep in all areas where the unsuccessful predominate is a type of gerrymandering which vitiates a significance test. Hundreds of such irregular lines might be drawn. Therefore, it would be expected that in any sample some line could be found yielding a difference "significant" at the 1 per cent level. At best, the irregular line sets up a hypothesis which, if found to yield a significant difference in a new and independent sample, could be taken as possibly true.

The law of parsimony enters this problem. Wherever a set of data may be explained equally well by two hypotheses, it is sound practice to accept the simpler hypothesis. Irregular cutting lines, and explanations in terms of patterns of scores, are sometimes justified and necessary. But in this case the difference between the groups is explained as well by the hypothesis that the successful boys give more  $M$ 's as by any non-spurious test of the  $W:M$  relationship. Therefore, procedure 5 is the soundest expression of the significance of the Margulies data. With more cases, this difference might be found to be truly significant.

In the above analysis, we find again that different procedures, more than one of which is mathematically sound, give different conclusions. The results from chi-square are less compatible with the null hypothesis than is the critical ratio. Chi-square applied to a dichotomy gives evidence of a possible relationship whereas chi-square applied to the frequency distribution does not. Attention is again drawn to the necessity of regarding with great suspicion any significance test based on a complex hypothesis set up to take advantage of the fluctuations of frequencies in a particular sample. Finally, it is noted that explanations in terms of ratios and patterns should not be sought unless they can account for observed differences more completely than can hypotheses in terms of single scores.

### *Treating Patterns of Scores*

Rorschach workers continually stress the importance of considering any score in relation to the unique pattern of scores for the individual. While this is done in clinical practice, there is no practical statistical procedure for studying the infinite complex interrelations of scores and indications on which the clinician relies. Instead of considering the individual patterns, the



statistician can at best study certain specific patterns likely to occur in many records. A pattern can be exceedingly complex; there is no statistical reason to prevent one from studying whether (for example) more men than women show high-*S*-on-colored-cards-accompanied-by-emphasis-on-*M*-and-excess-of-*CF*-over-*C*. The only limitation the statistical approach imposes is that the same pattern of scores must be studied in all cases.

Patterns of scores may be considered by means of composite scores, by definition of significant "signs," and by the pattern-tabulation method. The composite score is simply an attempt to express, in a formula, some psychologically important relationship. Examples include the *M*: *sum C* ratio, and the more complex composites developed by Hertz or Rapaport. These scores may be treated statistically like any score on a single category, although most of them are ratios or differences and suffer from the limitations already discussed.

*Comparing incidence of "signs."* The "signs" approach has been widely used. It is simple and well-adapted to the Rorschach test. Normally, an investigator identifies some characteristic of a special group, such as neurotics, from clinical observation. Then this characteristic is defined in a sign, i.e., a rule for separating those having the characteristic. One such sign, for example, is  $FM > M$ . After the investigator hypothesizes that some sign is discriminative, the necessity arises for making a test of significance to see if the sign is found more often in the type of person in question. One may soundly compare a new sample of the diagnosed group with a control sample by noting the frequency of the sign in each group and applying chi-square. This procedure is illustrated in studies by Hertzman and Margulies (38), and Ross (50).

The investigator may invent his own signs, if he follows due precautions to avoid misleading inflation of probabilities. Often it is easier and equally wise to use a predetermined set of signs. The most useful set of signs available at present is the Munroe check list. She has identified numerous ratios and patterns of scores which she considers significant of disturbance in her subjects (adolescent girls). She has stated that she does not think of her method as a set of signs (41), but the difference between her list and others appears to be (a) that it provides an inclusive survey of all deviations in a record and (b) that the list is designed as a whole to minimize duplication from sign to sign. There is no reason why two groups may not be compared by applying the check list to every record, and then comparing the groups on the frequency with which they receive each of the possible checks. Chi-square is the proper significance test, as used in one of Munroe's studies (43). The Munroe signs sometimes are simply defined (e.g. *P*— is 0 or 1 popular

response), but some involve patterns of several scores (thus the sign  $FM+$  is defined in terms of  $FM$ ,  $M$ , and  $R$ ).

*Pattern tabulation.* Pattern tabulation is a method devised by Cronbach for the study of relations between two or three scores (10). It has the advantage of permitting one to study the distribution of patterns in a group. To deal with any set of three scores, e.g.  $W$ ,  $D$ ,  $Dd$ , one normalizes the three scores for each person, and considers the resulting profile. The profile is expressed numerically in terms of the deviation of the converted scores from their average for each person. These three scores can be plotted on a plane surface, and the resulting scattergram shows the distribution of patterns in a group. If two groups are compared, any type of pattern found more commonly in one group than another can be identified, and the difference in frequency tested by chi-square. The significance level for rejecting the null hypothesis must be set conservatively, as this method involves many implied significance tests. An analysis of variance solution is also possible but not recommended in view of the fact that distributions of patterns are often non-normal.

This method cannot consider hypotheses involving more than three scores at once. It functions best when the three scores are equally reliable and equally intercorrelated. It encounters difficulty due to the fact that some Rorschach scores are unreliable, since any serious error of measurement in one score throws an error into the profile. The method does, however, appear flexible and especially useful for such meaningful patterns as  $W-D-Dd$  and  $M-sumC-F$ .

Another group of procedures leading to composite formulas for discriminating groups is treated in the text section.

### *Discrimination by Composite Scores*

In many problems, it is desired to use the Rorschach to discriminate between two groups. Thus one might seek a scoring formula to predict pilot success, or a "neurotic index" to screen neurotics from a general population. The methods used to arrive at composite scores are the check list, the multiple regression equation, and the discriminant function.

*Check list scores.* The check list consists of a set of signs. Each person is scored on the check list and the total number of signs or checks is taken as a composite score. This method has had considerable success, notably in Munroe's study (42) and in the formula of Harrower-Erickson and Miale for identifying insecure persons. There are no serious statistical problems in the use of check lists. The total score can be correlated (though eta may be



preferable to  $r$ ). Differences between groups may be tested for significance, preferably by chi-square. Chi-square is advised because a difference in the non-deviate range is rarely psychologically significant; the investigator is usually concerned with the proportion of any group in the deviate range. Buhler and Lefever justifiably applied analysis of variance to their check list score, to study its ability to differentiate clinical groups (5).

Problems do arise, however, in developing check list scores. A common method is to compare two groups on one raw score after another, noting where their means differ. Each score where a difference arises is then listed as a sign, and counted positively or negatively in obtaining the check list score for each case. This method takes advantage of whatever differences between samples arise just from accidents of sampling. If sample A exceeds B in mean  $M$ , allowing one point in the total score for high  $M$  will help discriminate A's and B's. In this sample, the A's will tend to earn higher check list scores. But often in a new sample such a difference will not be confirmed, and the  $M$  entry in the composite will not discriminate.

One study employing the sign approach should be pointed out to Rorschach workers. Davidson (12) sought to determine the relationship between economic background and Rorschach performance in a group of highly intelligent children. Her treatment of data is noteworthy because of her procedures; statistics are applied with great intelligence, new procedures being adopted for each new type of comparison. While the reviewer disagrees with some of the judgments she made in selecting procedures, her treatment is free from overt errors and well worth study by other Rorschach investigators.

Davidson divided her 102 cases among seven economic levels. She studied the Rorschach performance in various ways. First, she made a clinical analysis of each child, and placed him in one of nine categories (introvert adjusted, childish, constricted, disturbed, etc.). The distribution which resulted is a  $7 \times 9$  table. Recognizing that the expected frequency in each cell is quite small, she combined groups to form a  $3 \times 3$  table before applying the chi-square test for significance. This same type of condensation would have been advisable in some other comparisons she made, such as that between personality pattern and IQ. Davidson next applied a list of signs, and obtained for each case the total number of signs of maladjustment. The number of signs was correlated with economic level, and the correlation was shown not to differ significantly from zero. She tested the significance of the difference in mean number of signs by the critical ratio. These procedures appear well suited to her data. A third attack on the data treats one Rorschach score at a time. Here Davidson placed her cases in seven categories, ranging from highest to lowest economic level. By analysis of variance, she

demonstrated that differences among the seven groups were significant only for a few of the scores. The application of analysis of variance to continuous data appears to have been an unwise decision. Analysis of variance, like chi-square or eta applied to a variable divided in several categories, ignores the order of the categories. Consider the following set of means in the score  $M\text{--}sumC$ :

Economic level	1	2	3	4	5	6	7	Total
Mean score	1.17	1.86	1.29	0.96	-0.75	-0.13	-0.71	0.63

The downward trend from Group 1 to Group 7 gives great support to the hypothesis that this score is related to economic level. Analysis of variance estimates significance without considering this trend; the same significance estimate would be arrived at if Group 2 had had the mean of -0.13 and Group 6 the mean of 1.86. Davidson might have computed the correlation between each score and the economic level, but the skewness of some Rorschach scores weighs against this suggestion. The simplest procedure for testing this trend is to split the group into a  $2 \times 2$  table by combining adjoining categories in the economic scale, and dichotomizing the Rorschach score at a convenient point. Chi-square would then give the significance estimate. Such a procedure might have yielded significant differences in several instances where Davidson found none.

In justice to Davidson, it should be repeated that her data have been singled out for critical comment because of their exactness and completeness, rather than because they were improperly handled. The foregoing suggestions point to ways in which she might have arrived at additional important findings.

*The multiple regression formula.* A limitation of check lists is that they are simple additive combinations of signs which individually discriminate. But in such a composite a given trait may enter several times if it is reflected in several signs, and thus have greater proportionate weight than it deserves. The check list method does not allow for the possibility that certain signs may reinforce each other to indicate more severe maladjustment than is indicated by a combination of two other non-reinforcing signs, or for the possibility that two signs which are individually unfavorable may operate to neutralize each other. Multiple regression and the discriminant function are more powerful procedures than the usual check list score, because they consider the intercorrelations of scores and weight them accordingly.

By multiple correlation, one arrives at a regression equation which assigns weights to those variables which are correlated with a criterion and relatively uncorrelated with each other. This formula may be used to pre-



dict or to discriminate between groups. One such formula is that of the Air Force, used in its attempt to predict pilot success (21):

$$2(Dd + S\%) + 6FM + 8W - 1.5D\% + R - (VIII - X\%).$$

Multiple correlation does not seem especially promising for Rorschach studies. Even such an elaborate formula as that above turns out to have little or no predictive value when applied to a fresh sample. Even if it were stable, any formula of this type must assume that strength in one component compensates linearly for weakness in another. In this formula, emphasis on *Dd* would cancel weakness in *FM*, in estimating a man's pilot aptitude. It is most unlikely that the factors cancel each other in the personality itself. The simple linear regression formula provides an efficient weighting if the assumption of linear compensation is valid, but interrelations between aspects of personality are probably far too complex to be adequately represented in this way. The most that can be said for a regression formula is that, when derived on large samples (and this may require 5000 cases), it is a more precise prediction formula than the simple check list score can be. It cannot hope to yield very accurate predictions if interrelations within personality are as complex as Rorschach interpreters claim.

The discriminant function is a relatively new technique giving a formula which will separate two categories of men as thoroughly as possible from a mixed sample. It would be used to develop an effective index for separating good from poor pilots (not for predicting which man will be best, as the regression formula does) or for distinguishing organics and feeble-minded. A practical procedure for dealing with multiple scores has just been published by Penrose (45), and has not been employed in the Rorschach research. It appears likely to have real value in studies comparing different types of subjects.

Like the regression formula, however, the discriminant function provides a set formula. In this formula, it is assumed that one factor compensates for or reinforces weakness in another factor. The interactions within personality are probably too complex to be fully expressed by linear or quadratic discriminant functions.

### *Correlation and Reliability*

*Correlations of scores.* For one purpose or another several studies have tried to show the relationship between the several Rorschach scores or between Rorschach scores and external variables. The conventional procedure

for showing that two characteristics are associated is to compute a product-moment correlation between the variables. This has been done by Kaback (29), Vaughn and Krug (64), and others.

This method is unable to show the full relationship between variables when the regression of one on the other is curvilinear. Such a regression often occurs when one variable or both have a sharply skewed distribution. In fact, Vaughn and Krug note that one of their plots is curvilinear. The extent to which association may be underestimated is suggested by the following data. The data used are taken from tests administered individually by Audrey Rieger to several hundred applicants for employment, usually for managerial or technical positions. The tests were carefully scored by the Beck method. Generalization from the data must be limited because the group is not a sample of any clearly defined population. For 268 men, the product-moment correlation between  $D$  and  $Dd$  is .735. The curvilinear correlations are  $\eta_{DdD}$ , .785;  $\eta_{DDd}$ , .823. There is significant curvilinearity. If  $D$  and  $Dd$  are normalized, the regression becomes linear except for the effect of tied scores where  $Dd = 0$ ; for the converted scores,  $r = .767$ .

Brower employs rank-difference correlations in comparing certain Rorschach scores to physiological measures (3). This is a useful method for small samples and is equally sound for linear and non-linear regressions. Thus, a rank-correlation of  $W/M$  with another score is the same except for sign as the correlation for the inverted ratio  $M/W$ , but the product-moment correlations are far different.

The rank method does have the disadvantage of weighting heavily the small and unreliable differences in the shorter end of skew distributions, where many cases have the same rank. This might lower the correlations for a score like  $Fc$ , but is not a difficulty with scores distributed more symmetrically over a wide range, such as  $F$  or  $VIII-X\%$ . Normalizing has the same disadvantage. This is a reflection of the inability of the test to discriminate finely among cases in the modal end of a severely skewed distribution.

*Reliability coefficients.* Test reliability is ordinarily estimated by the retest or the split-half method. These methods are not very appropriate for the Rorschach test, the former because of memory from trial to trial, the latter because the test cannot be split into similar halves. Nevertheless, both methods have been used in the absence of better procedures.

The split-half method introduces a statistical problem which not all investigators have noted, namely, that the Spearman-Brown formula must not be applied to ratios with variable denominators such as  $W\%$  and  $M/\text{sum}C$ . Methods for estimating the reliability of ratio scores have been treated elsewhere (7, 8), but these procedures are not useful when the denominator is relatively unreliable (as in  $M/\text{sum}C$ ).



It is desirable to estimate reliability of scores separately for records of varying length. Vernon (65) found that Rorschach scores were much more reliable for cases where  $R > 30$  than when  $R < 30$ . This implies that it is unsatisfactory to estimate just one reliability coefficient for a group with varied  $R$ . Instead, the standard error of measurement of  $W$ , or  $W\%$ , should be determined separately for cases where  $R = 10-15$ ,  $R = 15-25$ ,  $R = 25-35$ , or some such grouping.

The reliability of patterns of scores is a difficult problem. If both  $M$  and  $W$  were perfectly reliable, any pattern or combination based on the two scores would also be perfectly reliable. But these scores are unstable; subjects vary from trial to trial in  $M$  or  $W$  or both. Nevertheless, Rorschach users insist that the "pattern" of scores is stable. If there is any substance to this claim, it means that certain definable configurations of the scores are stable even though the separate scores are not. The configurations may be as simple as the  $W/M$  ratio or may be complex structures of several scores. One may establish the reliability of any composite score by obtaining two separate estimates from independent trials of the test.

The method of determining reliability by independent estimates has rarely been used. A study by Kelley, Margulies, and Barrera (30) is of interest, even though based on only twelve cases. The Rorschach was given twice, and between the trials a single electroshock was given, reportedly sufficient to wipe out memory of the first trial without altering the personality. In the records so obtained,  $R$  shifted as much as 50 per cent from trial to trial, and absolute values of some other scores shifted also. In several cases where scores shifted, it can be argued that the *relationship* between the scores did not shift and that the two records would lead to similar diagnoses. The authors made no attempt at statistical treatment. Probably this ingenious procedure will rarely be repeated. Useful studies could certainly be made, however, by comparing performance on two sets of ink blots without shock (cf. Swift, 57). Even if the two sets are not strictly equivalent, the data would indicate more about the stability of performance than any methods so far employed.

At first glance, it appears logical to set up composite scores, obtain two separate estimates, and correlate them. Even this is unsuitable for Rorschach problems, however. As pointed out before, a given ratio such as 20%  $W$  or  $W/M$  2.0 has different meaning in different records, depending on the absolute value of  $W$ . The pattern might conceivably be defined by a curvilinear equation, but this becomes unmanageable, especially as several variables enter a single pattern. The problem is one of defining when two patterns are psychologically similar, and of defining the magnitude of the difference

when they are not equivalent. No one would contend that the  $W/M$  balance is unchanged if a subject shifts from  $12 W : 2M$  to  $60 W : 10M$ . The problem is to define and measure the balance in a numerical way. The approach pattern  $W-D-Dd$  has three dimensions. If we wish to estimate reliability by comparing two sets of these three scores we have a six-dimensional array, for which no present methods are adequate. So far, even the pattern-tabulation method reduces such data to only four dimensions, which leaves the problem still unmanageable. All that can be recommended is that additional attention be given to this challenging problem. We can now obtain adequate evidence on the stability of Rorschach patterns only by such a method as Troup's (62), discussed in the first section of this paper. It will be recalled that she had two sets of records interpreted clinically, and employed blind-matching to show that the inferences from the Rorschach remained stable.

Two unique but entirely unsound studies of Fosberg (15, 16) employed a novel procedure to estimate the reliability of the total pattern. He gave the test four times, under varied directions. He then compared the four records for each person. In one study he used chi-square to show that the psychograms for each person corresponded. But this statistical test merely showed that the  $D$  score in record 1 is nearer to  $D$  in record 2 than it is to  $W$ ,  $C$ , or other scores. That is, he showed that the scores were not paired at random. But, since each score has a relatively limited range for all people—i.e.,  $D$  tends to be large,  $m$  tends to be small, etc.—he would have also obtained a significantly large chi-square if he had applied the same procedure to four records from *different* persons. One may also point out that finding a  $P$  of .90 does not prove that two records do come from the same person, but only that the null hypothesis is tenable, or possibly true. Fosberg's second study, using correlation technique, is no sounder than the first. Here the two sets of scores for one person were correlated. That is, pairs of values such as  $W_1 - W_2$ ,  $D_1 - D_2$ , etc. were entered in the same correlation chart. As before, the generally greater magnitude of  $-D$  causes the two sets to correlate, but high correlations would have been obtained if the scores correlated came from two different subjects.

Objection must also be made to several procedures and inferences of Buhler and Lefever (5), in their attempts to demonstrate the dependability of their proposed Basic Rorschach Score. 1) They used the split-half method on the total score, by placing half the signs in one list, the other half in a second list, and scoring each person on both lists (5, p. 112). They then correlated the two halves to indicate reliability. Because the correlation was computed on cases used to determine the scoring weights for the items, the resulting correlation is spuriously high. Even if new cases were obtained, the



split-half method would be incorrect because the check list items are not experimentally independent. A single type of performance enters into a great number of separately scored signs (in their check list, *M* affects items 1, 2, 5, 6, 7, 8, 10, 11, 12, 51, 52, 53, 86, 93, 94, 95, 96, 99, 100, 101, and 102). A "chance" variation in *M* would alter the score on all these categories, and would spuriously raise the correlation unless these linked categories were concentrated in the same half of the test. 2) They derived separate sets of weights from the comparison of Normals vs. Schizophrenics, Nurses vs. Schizophrenics, and other groups. The correlation between the scoring weights is high, which they take as evidence for reliability (pp. 112 ff.). At least one serious objection is that the weights were derived in part from the same cases. If, by sampling alone, *FK* happened to be rare among the Schizophrenic group, this would cause the sign *FK* to have a weight in both the Normal-Schizophrenic key and the Nurse-Schizophrenic key. The evidence is not adequate to show that the weights would be the same if the two keys were independently derived. This objection does not apply to another comparison of the same general type, where the four samples involved had no overlap. 3) Certain papers were scored repeatedly, using sets of weights derived in comparable but slightly different ways (p. 116). The correlations of the resulting sets of scores are advanced as evidence of reliability. Any correlation of separate scorings of the same set of responses is in part spurious. If responses of individual subjects were determined solely by chance, there would still be a correlation when keys having any similarity to each other were applied to the papers. The reliability of the performance of the subject, and that is what reliability coefficients are supposed to report, cannot be revealed by rescorings of the same performance.

### *Conclusions*

The foregoing analysis and the appended bibliography are convincing evidence that Rorschach workers have sought statistical confirmation for their hypotheses. But the analysis also shows that the studies have been open to errors of two types: 1) erroneous procedures have led to claims of significance and interpretations which were unwarranted; and 2) failure to apply the most incisive statistical tests has led workers to reject significant relationships. So widespread are errors and unhappy choices of statistical procedures that few of the conclusions from statistical studies of the Rorschach test can be trusted. A few workers have been consistently sound in their statistical approach. But some of the most extensive studies and some

of the most widely cited are riddled with fallacy. If these studies are to form part of the base for psychological science, the data must be reinterpreted. Perhaps 90 per cent of the conclusions so far published as a result of statistical Rorschach studies are unsubstantiated—not necessarily false, but based on unsound analysis.

Few of the errors were obvious violations of statistical rules. The Rorschach test is unlike conventional instruments and introduces problems not ordinarily encountered. Moreover, statistical methods for such tests have not been fully developed (11). It is most important that research workers using the Rorschach secure the best possible statistical guidance, and that editors and readers scrutinize studies of the test with great care. But statisticians have a responsibility too, to examine the logic of Rorschach research and the peculiar character of clinical tests, in order to sense the limitations of conventional and mathematically sound procedures.

Present statistical tools are imperfect. And no procedure is equally advisable for all studies. Within these limitations, this review has suggested the following guides to future practice.

1. Matching procedures in which a clinical synthesis of each Rorschach record is compared with a criterion are especially appropriate.
2. If ratings are to be treated statistically, it is often advisable to dichotomize the rating and apply chi-square or bi-serial  $r$ .
3. Common errors which must be avoided in significance tests are:
  - a. Use of critical ratio and uncorrected chi-square for unsuitably small samples.
  - b. Use of sample values in the formula for differences between proportions.
  - c. Use of formulas for independent samples when matched samples are compared.
  - d. Interpretation of  $P$ -values without regard for the inflation of probabilities when hundreds of significance tests are made or implicitly discarded.
  - e. Acceptance of conclusions when a significant difference is found with a hypothesis based on fluctuations in a particular sample.
4. Counting procedures are in general preferable to additive methods for Rorschach data. The most widely useful procedures are chi-square and analysis of differences in mean rank. These yield results which are invariant when scores are transformed.
5. Normalizing scores is frequently desirable before making significance tests involving variance.



6. Where groups differ in total number of responses, this factor must be held constant before other differences can be soundly interpreted. Three devices for doing this are: rescoring a fixed number of responses on all papers, constructing subgroups equated on the number of responses, and analyzing profiles of normalized scores (pattern tabulation).

7. Ratio and difference scores should rarely be used as a basis for statistical analysis. Instead, patterns should be defined and statistical comparisons made of the frequency of a certain pattern in each group. Use of chi-square with frequencies of Rorschach "signs" is recommended.

8. Multiple repression and linear discriminant functions are unlikely to reveal the relationships of Rorschach scores with other variables, since the assumption of linear compensation is contrary to the test theory.

9. Rank correlation, curvilinear correlation, or correlation of normalized scores are often more suitable than product-moment correlation.

10. No entirely suitable method for estimating Rorschach reliability now exists. Studies in the area are much needed.

There are in the Rorschach literature numerous encouraging bits of evidence. The question whether the test has any merit seems adequately answered in the affirmative by studies like those of Troup, Judith Krugman, Williams (69), and Munroe. Supplemented as these are by the testimony of intelligent clinical users of the test, there is every reason to treat the test with respect. One cannot attack the test merely because most Rorschach hypotheses are still in a pre-research stage. Some of the studies which failed to find relationships might have supported Rorschach theory if the analysis had been more nearly perfect. How accurate the test is, how particular combinations of scores are to be interpreted, and how to use Rorschach data in making predictions about groups are problems worth considerable effort. With improvements in projective tests, in personality theory, and in the statistical procedures for verifying that theory, we can look forward to impressive dividends.

## BIBLIOGRAPHY

1. ABEL, T. M. Group Rorschach testing in a vocational high school. *Rorschach Res. Exch.*, 1945, 9, 178-188.
2. BECK, S. J. Personality structure in schizophrenia. *Nerv. and Ment. Dis. Monogr.*, 1938, No. 63.

3. BROWER, D. The relation between certain Rorschach factors and cardiovascular activity before and after visuo-motor conflict. *J. Gen. Psychol.*, 1947, 37, 93-95.
4. BROWN, R. R. The effect of morphine upon the Rorschach pattern in post-addicts. *Amer. J. Orthopsychiat.*, 1943, 13, 339-342.
5. BUHLER, C., BUHLER, K., & LEFEVER, D. W. *Rorschach standardization studies. Number I. development of the basic Rorschach score.* Los Angeles: C. Buhler, 1948.
6. COCHRAN, W. G. The chi-square correction for continuity. *Iowa St. Col. J. Sci.*, 1942, 16, 421-436.
7. CRONBACH, L. J. The reliability of ratio scores. *Educ. Psychol. Msmt.*, 1941, 1, 269-278.
8. CRONBACH, L. J. Note on the reliability of ratio scores. *Educ. Psychol. Msmt.*, 1943, 3, 67-70.
9. CRONBACH, L. J. A validation design for personality study. *J. Consult. Psychol.*, 1948, 12, 365-374.
- ✓ 10. CRONBACH, L. J. Pattern tabulation: a statistical method for treatment of limited patterns of scores with particular reference to the Rorschach test. *Educ. Psychol. Msmt.*, 1949, 9, 149-171.
11. CRONBACH, L. J. Statistical methods for multi-score tests. Paper presented before the Biometrics Section, American Statistical Association, December, 1948. *J. Clin. Psychol.*, 1950, 6, 21-25.
12. DAVIDSON, HELEN H. *Personality and economic background.* New York: King's Crown Press, 1945.
13. EDWARDS, A. L. Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika*, 1948, 13, 185-187.
- ✓ 14. FESTINGER, L. The significance of difference between means without reference to the frequency distribution function. *Psychometrika*, 1946, 11, 97-105.
15. FOSBERG, I. A. Rorschach reactions under varied instructions. *Rorschach Res. Exch.*, 1938., 3, 12-31.
16. FOSBERG, I. A. An experimental study of the reliability of the Rorschach technique. *Rorschach Res. Exch.*, 1941, 5, 72-84.
17. FREEMAN, H., RODNICK, E. H., SHAKOW, D., & LEBEAUX, T. The carbohydrate tolerance of mentally disturbed soldiers. *Psychosom. Med.*, 1944, 6, 311-317.
18. GANN, E. *Reading difficulty and personality organization.* New York: King's Crown Press, 1945.
19. GOLDFARB, W. A. A definition and validation of obsessional trends in the Rorschach examination of adolescents. *Rorschach Res. Exch.*, 1943, 7, 81-108.
20. GOLDFARB, W. Effects of early institutional care on adolescent personality. *Amer. J. Orthopsychiat.*, 1944, 14, 441-447.
21. GUILFORD, J. P. (Ed.) *Printed classification tests.* AAF Aviation Psychology Program Research Reports, No. 3. Washington: Government Printing Office, 1947.



22. GUSTAV, ALICE. Estimation of Rorschach scoring categories by means of an objective inventory. *J. Psychol.*, 1946, 22, 253-260.
23. HARRIS, R. E., & CHRISTIANSEN, C. Prediction of response to brief psychotherapy. *J. Psychol.*, 1946, 21, 269-284.
24. HARRIS, T. M. The use of projective techniques in industrial selection. In *Exploring individual differences*, American Council on Education Studies, Series 1, No. 32, 1948. Pp. 43-51.
25. HERTZ, MARGUERITE R. Personality patterns in adolescence as portrayed by the Rorschach ink-blot method: I. The movement factors. *J. Gen. Psychol.*, 1942, 27, 119-188.
26. HERTZMAN, M. A comparison of the individual and group Rorschach tests. *Rorschach Res. Exch.*, 1942, 6, 89-108.
27. HERTZMAN, M., & MARGULIES, H. Developmental changes as reflected in Rorschach test responses. *J. Genet. Psychol.*, 1943, 62, 189-215.
28. HERTZMAN, M., ORLANSKY, J., & SEITZ, C. P. Personality organization and anoxia tolerance. *Psychosom. Med.*, 1944, 6, 317-331.
29. KABACK, G. R. *Vocational personalities: an application of the Rorschach group method*. New York: Bureau of Publications, Teachers Coll., Columbia Univ., 1946.
30. KELLEY, D. M., MARGULIES, H., & BARRERA, S. A. The stability of the Rorschach method as demonstrated in electric convulsive therapy cases. *Rorschach Res. Exch.*, 5, 1941, 35-43.
31. KRUGMAN, J. I. A clinical validation of the Rorschach with problem children. *Rorschach Res. Exch.*, 1942, 6, 61-70.
32. KRUGMAN, M. Psychosomatic study of fifty stuttering children. *Amer. J. Orthopsychiat.*, 1946, 16, 127-133.
33. KURTZ, A. K. A research test of the Rorschach test. *Personnel Psychol.*, 1948, 1, 41-51.
34. LEVERETT, H. M. Table of mean deviates for various portions of the unit normal distribution. *Psychometrika*, 1947, 12, 141-152.
35. LINDQUIST, E. F. *A first course in statistics* (Revised ed.). Boston: Houghton Mifflin, 1942.
36. MCCANDLESS, B. R. The Rorschach as a predictor of academic success. *J. Appl. Psychol.*, 1949, 33, 43-50.
37. MCNEMAR, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 1947, 12, 153-157.
38. MARGULIES, H. Rorschach responses of successful and unsuccessful students. *Arch. Psychol., N. Y.*, No. 271. New York, 1942.
39. MELTZER, H. Personality differences between stuttering and non-stuttering children. *J. Psychol.*, 1944, 17, 39-59.
40. MONTALTO, F. D. An application of the Group Rorschach technique to the problem of achievement in college. *J. Clin. Psychol.*, 1946, 2, 254-260.
41. MUNROE, RUTH L. Objective methods and the Rorschach blots. *Rorschach Res. Exch.*, 1945, 9, 59-73.

42. MUNROE, RUTH L. Prediction of the adjustment and academic performance of college students by a modification of the Rorschach method. *Appl. Psychol. Monogr.*, 1945, No. 7.
43. MUNROE, RUTH L. Rorschach findings on college students showing different constellations of subscores on the A.C.E. *J. Consult. Psychol.*, 1946, 10, 301-316.
44. PEATMAN, J. G. *Descriptive and sampling statistics*. New York: Harper, 1947.
45. PENROSE, L. S. Some notes on discrimination. *Ann. Eugenics*, 1947, 13, 228-237.
46. PIOTROWSKI, Z., CANDEE, B., BALINSKY, B., HOLTZBERG, S., & VON ARNOLD, B. Rorschach signs in the selection of outstanding young male mechanical workers. *J. Psychol.*, 1944, 18, 131-150.
47. RAPAPORT, D. *Diagnostic psychological testing, Vol. II*. Chicago: Year Book Publishers, 1946.
48. RICHARDSON, L. H. The personality of stutterers. *Psychol. Monogr.*, 1944, 56, No. 7.
49. RICKERS-OVSIANKINA, M. The Rorschach test as applied to normal and schizophrenic subjects. *Brit. J. Med. Psychol.*, 1938, 17, 227-257.
50. ROSS, W. D. The contribution of the Rorschach method to clinical diagnosis. *J. Ment. Sci.*, 1941, 87, 331-348.
51. ROSS, W. D., FERGUSON, G. A., & CHALKE, F. C. R. The Group Rorschach Test in officer selection. *Bull. Canad. Psychol. Assn.*, 1945, 84-86.
52. ROSS, W. D., & ROSS, S. Some Rorschach ratings of clinical value. *Rorschach Res. Exch.*, 1944, 8, 1-9.
53. SARBIN, T. R., & MADOW, L. W. Predicting the depth of hypnosis by means of the Rorschach test. *Amer. J. Orthopsychiat.*, 1942, 12, 268-271.
54. SCHMIDT, H. O. Test profiles as a diagnostic aid: the Rorschach. *J. Clin. Psychol.*, 1945, 1, 222-227.
55. SIEGEL, M. G. The diagnostic and prognostic validity of the Rorschach test in a child guidance clinic. *Amer. J. Orthopsychiat.*, 1948, 18, 119-133.
56. SNEDECOR, G. W. *Statistical methods*. Ames, Iowa: Iowa State College Press, 1940.
57. SWIFT, J. W. Reliability of Rorschach scoring categories with preschool children. *Child Developm.*, 1944, 15, 207-216.
58. SWIFT, J. W. Rorschach responses of 82 pre-school children. *Rorschach Res. Exch.*, 1945, 9, 74-84.
59. SWINEFORD, F. A table for estimating the signature of the difference between correlated percentages. *Psychometrika*, 1948, 13, 23-25.
60. THOMPSON, G. M. College grades and the Group Rorschach. *J. Appl. Psychol.*, 1948, 32, 398-407.
61. THORNTON, G. R., & GUILFORD, J. P. The reliability and meaning of Erlebnistypus scores on the Rorschach test. *J. Abnorm. Soc. Psychol.*, 1936, 31, 324-330.
62. TROUP, E. A comparative study by means of the Rorschach method of personality development in twenty pairs of identical twins. *Genet. Psychol. Monogr.*, 1938, 20, 461-556.
63. TULCHIN, S., & LEVY, D. Rorschach test differences in a group of Spanish and English refugee children. *Amer. J. Orthopsychiat.*, 1945, 15, 361-368.



64. VAUGHN, J., & KRUG, O. The analytic character of the Rorschach inkblot test. *Amer. J. Orthopsychiat.*, 1938, 8, 220-229.
65. VERNON, P. E. The Rorschach inkblot test. *Brit. J. Med. Psychol.*, 1933, 13, 179-200.
66. VERNON, P. E. The matching method applied to investigations of personality. *Psychol. Bull.*, 1936, 33, 149-177.
67. WALKER, HELEN M. *Elementary statistical methods*. New York: Holt, 1943.
68. WERNER, H. Perceptual behavior of brain-injured, mentally defective children. *Genet. Psychol. Monogr.*, 1945, 31, 51-110.
69. WILLIAMS, M. An experimental study of intellectual control under stress and associated Rorschach factors. *J. Consult. Psychol.*, 1947, 11, 21-29.
70. YATES, F. The analysis of contingency tables with groupings based on quantitative characters. *Biometrika*, 1948, 35, 176-181.

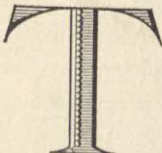
# VII

## *Summary*



*Marguerite R. Hertz*

*CURRENT PROBLEMS IN  
RORSCHACH THEORY AND  
TECHNIQUE<sup>1</sup>*

 THE LAST fifty years have been marked by rapid developments in the field of projective techniques. Because of their early promise as tools for penetrating the total personality and understanding it in all its dynamic interrelationships, more and more clinicians have included projective methods in their examinations and growing numbers of investigators have turned their attention to research in this area. Nevertheless, projective techniques are only in the beginning stages of their development. They have yet to be established as reliable and valid techniques.

The Rorschach method, on which I would like to concentrate, is the earliest historically and the most highly developed of all the projective techniques. Yet it still lacks unqualified scientific status. Since its earliest days, we have sought to increase its objectivity and to validate its hypotheses. We have applied statistical procedures and have broadened our application.

Reprinted from *J. Proj. Tech.*, 1951, 15, 307-338, by permission of the publisher and the author.

1. Delivered in part at the Symposium on Current Problems in the field of Projective Techniques, May 4, 1950, Midwestern Psychological Association, Detroit, Michigan.

The number of research studies published offers impressive evidence of the quickened tempo and broadened scope of Rorschach activity and gives some indication of its importance and acceptance in the fields of clinical psychology, anthropology and the social sciences. Nevertheless, many scientists (25, 33, 50, 141, 170) remain highly critical of these efforts because of limitations inherent in the method. They charge us with uncritical acceptance of data at face value, extreme subjectivity, and dependence on insights in the absence of detailed study and research. They censure what they call our disregard of the accepted rules of scientific verification, and grow impatient with our inability to show absolute and objective verification of our contentions. They are highly critical of our failure to use statistics and where we use them they point out errors in our procedures.

Whether or not these criticisms are conclusive, they suggest real and challenging problems.

## *Theory*

We have been criticized for failure to develop a basic underlying theory for our method and for paying little attention to the theoretical and conceptual side of our problems. There can be little doubt that there is merit in these criticisms. Theoretical and systematic formulations are sadly lacking in much of our work.

As we know, Rorschach himself developed no new discipline. His thinking was influenced by the development of his time in many fields—faculty psychology, Gestalt psychology, personalistic psychology, typology, psychiatry, psychoanalysis, and psychodiagnosis. His concentration was on *method*, a method to study the personality as a functioning whole. He formulated no specific theory of personality.

Despite the fact that Rorschach's monograph appeared as early as 1921, few systematic attempts have been made to get at these fundamental considerations. We have amassed fruitful hypotheses; we have accumulated facts which we have found valuable in clinical work; but so far, we are without a well defined theory to integrate either our hypotheses or our facts or both.

Those of us in the field have not, however, worked in a vacuum. Most of us have been influenced by psychoanalytic theory and/or by the various psychological theories which emphasize a global view of personality. A review of the Rorschach literature will reveal at once many constructs included in current theories variously called functional, holistic, dynamic, organismic and the like. Thus concepts of figure and ground, the role of the part as a function of the whole, systems of stress and field formations, em-



phasis on the unity of personality and the dynamic interrelationships among the various components of personality, emphasis on the organized character of all responses, all appear in some form in Rorschach data, have their background in Gestalt theory, and are related to some form of field theory. Again, Lewinian concepts of degree of differentiation, rigidity, harmonious and disharmonious structure, and the like appear frequently in Rorschach work. In like manner, the psychoanalytic treatment of the determinants of mental life, the use of such concepts as mechanisms of defense, symbolism, conscious and unconscious motivation, and more, have been incorporated into Rorschach interpretation.

Thus it must be recognized that the use of psychological and psychoanalytic concepts from various schools of thought implies in a measure an acceptance on the part of Rorschach workers of the theoretical principles upon which these systems are based. To this extent, then, we have theoretical frames of reference within which we work.

In addition, in the last few years, we note serious efforts to relate the Rorschach method to underlying theory in the field of perception. Recent studies on the dynamics of the perceptual process either use the Rorschach blots directly or bear directly on the Rorschach response. Experimental work on the various factors which influence perception, studies on value and needs by Bruner and Goodman (19) and McClelland and Atkinson (105), on symbolic values by Bruner and Postman (21), on tension and tension release as organizing factors in perception by Bruner and Postman (20), and on personal values as selective factors in perception by Postman, Bruner and McGinnies (126), all help illuminate the perceptual processes in the Rorschach.

Using the Rorschach blots directly to study the temporal characteristics of perception, Weisskopf (174) investigated the influence of the time factor on Rorschach performance. More recently Stein (160) using a tachistoscopic administration of the Rorschach cards studied personality factors involved in the temporal development of Rorschach responses and made important conclusions on the adaptive functions of perception in the Rorschach.

These are of course meager beginnings in the experimental study of the perceptual process in order to formulate some theory of personality. Whether the Rorschach blots are used or not in the experimental design, it should be borne in mind that the problem of the theoretical foundation of the Rorschach method is not uniquely one for the Rorschach worker *per se*. It is much more general and awaits the development of a basic theory underlying problems of personality in general and projective methods in particular. The Rorschach worker must join with others of course in the effort to develop that theory.

Thus, while it is true that Rorschach workers borrow many theoretical formulations from widely different methodological and conceptual disciplines, it is important to recognize that they are not indifferent to fundamental theory and are not working in a vacuum. More and more they strive toward a fundamental understanding of the Rorschach response. For the time being, in the absence of generally accepted concepts, hypotheses, and specific theories, they can only repeat with Macfarlane that "the first step in projective research should be an explicit statement of concepts used and an orientation with respect to theoretical biases." (100, p. 406)

### *Objectivity versus Subjectivity*

Turning to the perennial problem of objectivity, here too there are many dissonant notes. There can be little doubt that there has been tremendous progress in the objectification and standardization of the Rorschach method in the area of administration and scoring. Books and articles have been published with detailed instructions as to procedure [Beck (10), Bell (11), Bochner and Halpern (15), Klopfer and Kelley (89), Rapaport et al. (129)]. More uniform and efficient methods of recording and scoring responses and summarizing data have been developed [Hertz (65, 66)]. Hutt and Shor (78) have presented a systematic analysis of the probing which adds considerably to the objectivity of this phase of the administration.

There are some aspects of the procedure, however, which have been the subject of much concern. Despite considerable uniformity, there are some variations in administration which demand systematic study. These include sparse instructions versus more elaborate and detailed instructions, use of a trial blot, extensive inquiry or minimal inquiry, inquiry after each card or at the end of the test, and inquiry with or without the cards in evidence.

Several studies have discussed changes which take place in Rorschach reactions under varied instructions. Schachtel (149) has written in detail on the influence of subjective factors introduced by the Rorschach situation on Rorschach performance. The effect of social influence on Rorschach records was demonstrated by Kimble (86), on situational and attitudinal influences by Luchins (99). More recently, Gibby (46) has studied the influence of varied experimental sets upon Rorschach intellectual factors and Milton (110) on the human movement factor. In all these studies, the Rorschach test performance shows the effect of varied experimental "sets." Just so, Rorschach performance may also be affected by other more casual changes in the administration.

Differences in the nature of the inquiry conducted may likewise affect



the Rorschach performance. Most examiners restrict themselves to indirect and non-committal questions, carefully avoiding all suggestions as to Rorschach factors or specific content. Yet direct questioning is used by many examiners. Rapaport recommends the procedure of asking directly for certain information. He advises, for example, that in investigating for human movement, the following questions be asked,—“Was he doing anything?” “Anything you noticed about his posture?” (129, p. 209).

Buhler, Buhler and Lefever (22) in their study of the Basic Rorschach Score have deviated from the usual procedure to such an extent that there is considerable doubt in my mind as to the validity of the application of their results to records obtained when the test is administered in the more orthodox fashion. They restrict the number of responses to “three to five,” yet they permit some persons to continue and give more responses. They make specific suggestions to their subjects during the course of the test itself involving such matters as location, populars, and animal and human details. They change the inquiry by introducing the technique of alternative questioning, which is, for the most part, direct and highly suggestive questioning. Thus after an answer, “Two dogs” to Card II, they ask, “Do they seem to be standing or sitting or is it just the shape of the dogs you see?”

It may be that such deviations from the conventional procedure do not give a picture different from that which a more orthodox administrator would obtain. On the other hand, it may be that the results which they report reflect at least in part their method. Which if true, remains to be demonstrated. Until then, we are not ready to apply the Basic Scores to records administered in a more orthodox fashion. Further study to establish the validity of this application is required.

In general, systematic studies should determine the various influences or modifications in procedure on the resultant response record. Comparison of research findings is frequently impeded if not rendered impossible because of these variations in method. McFate and Orr (106), for example, point to the fact that the adolescent group in the Brush Developmental studies (59, 61, 62, 63, 71) gives much higher total numbers of responses than their group of similar make-up and age. This may well be explained in terms of differences in method of administration.

Not only the administration, but also the scoring has been developed to the point where most examiners utilize the same or approximately the same categories even though they do not always use the same symbols. This is not true of the FM (animal movement) and the m (movement in nature) which are not yet universally accepted in this country. It is of interest to note that European workers have incorporated these factors into their scoring systems

(69). Psychograms for systematic tabulations have been developed by Klopfer and Kelley (89) and Hertz (66).

Various scoring criteria have been subjected to further study. Beck (10) and Hertz (65) have introduced some changes in their respective lists of normal details and popular factors. Frequency Tables for use in scoring form level likewise have been revised by both Beck (10) and Hertz (65). These tables add considerably to the objectivity of the scoring. Recognizing the need for a more objective appraisal of the form quality, Klopfer and Davidson (88) have introduced a rating scale for study, stressing three form qualities—accuracy, specification, and organization. Beck (10) has introduced an organizational factor, "Z," Hertz (66) a similar factor, "g." Problems associated with the popular factor have been suggested by Hallowell (52), who discussed this factor in relation to the varying qualitative and quantitative criteria used for their determination in relation to age levels and different cultural backgrounds. Further consideration of these technical scoring problems will enhance the objectivity and reliability of the method.

In order to objectify the scoring even further, psychometric scales have been developed by Zubin, Chute and Veniar (191) to reduce subjectivity to a minimum and to provide for more exact quantification of the Rorschach. These scales have been further developed by Zubin and Young (192). Unfortunately, little work has been done with these scales to establish their greater merit over the traditional method of scoring.

Recently, the Rorschach method has been extended to include some type of content analysis. The suggestions of Lindner (98) and Rapaport et al. (129) as to the meanings of specific content, Goldstein and Rothmann (49) on the evaluation of physiognomic responses, and Goldfarb (47) on the significance of the animal symbol, all offer interesting leads for research. Analysis of the content of the responses of homosexuals has been presented by Bergmann (13), and Due and Wright (38) has stimulated considerable research in this area.

More systematic studies offering valuable data are reported by Wheeler (176) on the content in homosexual records and by Reitzell (130) on content in the responses of hysterics, homosexuals and alcoholics. Shaw (151) has identified "sex populars" on the basis of a study of the records of male subjects and recommends their inclusion as an extension of the probing aspect of the procedure. An elaboration of content analysis in terms of hostility and anxiety has been presented by Elizur (39). Hertzman and Pearce (75) also offer interesting leads as to the meaning of the human factor in the Rorschach response.

Continued study of the various aspects of the content of the Rorschach responses will no doubt furnish the Rorschach examiner with valuable



meanings of and procedure for studying various types of responses and will add considerably to the objectivity of the method.

While the administration and the scoring of the Rorschach record has attained a measure of objectivity which promises to increase, the interpretation of the record is still a highly subjective matter. It is because of this subjectivity that we have been so frequently censured.

Most Rorschach examiners take the intuitive configurational approach when they interpret a record, studying the various Rorschach scores as interrelated and interacting configurations. There is, however, an intermediate approach less important perhaps but none the less necessary, in which the individual is studied in reference to his group. This involves norms.

Unfortunately, norms present an area in which development commensurate with our needs is still lacking. Indeed, there are many Rorschach examiners who avoid this intermediate step, stressing the greater importance of studying the individual as a unit by himself, and censuring the normative approach as quantitative, static, and sterile. Of course, consciously or unconsciously, they too employ norms, subjective norms which they have amassed as a result of their own experience. As such, however, they are unreliable because they have not been systematically subjected to verification.

Granting that an individual must be studied as a unique personality, the interpreter must have some objective and reliable frame of reference to evaluate the functioning of his subject. He cannot understand him well unless he compares him with other personalities in his group. Time and time again we see interpretations of adolescent records, for example, in which the interpreter stresses personality deviations and psychoneurotic disturbances, sometimes even more serious pathology. Yet these very matters may be seen as not abnormal when judged by the normative material available. If we know something about the general characteristics of the adolescent personality, we have a perspective in which each adolescent may be placed. This is similarly true of other age groups and other developmental periods as well as other cultural groups.

As for Rorschach normative data, we know considerably more about what to expect of the "normal" individual on a Rorschach record than we did a few years ago. Recent studies have filled the gaps for different age levels. Children's norms based on adequate samples have been reported by Davidson (36), Ford (43), Hertz and Ebert (72), Kay and Vorhaus (84), Swift (167) and Vorhaus (171); adolescent norms by Hertz (59, 61, 62), Hertz and Baker (71), Hertzman and Margulies (74), Margulies (103), and McFate and Orr (106). Rabin and Beck (128) have presented norms for schoolchildren, ages six to twelve. Normative material for adults appear in the manuals of

Beck (10), Bockner and Halpern (15), Klopfer and Kelley (89), and Rapaport et al. (129).

Despite these studies, we are not without censure because of the inadequacy of norms at many age levels and because of the inadequate samples on which so many of them are based. Again, some of the studies of children are based on scoring criteria developed with adult records (Swift [167], Ford [43], and Stavrianos [159]). Their usefulness has therefore been challenged, and correctly so.

Again, comparison of published norms of similar age groups shows some discrepancies. Further study is needed to revise them or to explain the lack of agreement. A study of the discrepancies may well contribute to a better understanding of the nature of the factors involved.

The problem is not only to amass norms appropriate to the group but also to use and to interpret them wisely. Thus, it is elementary to emphasize that normative material should be expressed in terms of a range and that deviations be identified both above and below that range. Many studies fail to take into consideration this elementary fact, causing errors in the application of normative material. Thus Muench (114) in his study on the evaluation of psychotherapy, lists several so-called adjustment patterns viewing maladjustment in terms of only one deviation. He indicates that %A for example is expected to be no greater than 50 per cent. Anything below that figure, 15 per cent for example, he terms as satisfactory adjustment. This percentage falls considerably outside the range of normality and may not reflect adjustment at all. This error is repeated several times in his study.

The interpretation of scores in reference to a normative range is likewise important. While general meanings may be ascribed to certain norms, it is generally known that the meaning of a particular score in a particular case depends upon the interrelationship of that score with various other items in the record. We are told, for example, that two to three human movement responses may be expected in the record of a "normal" individual. The interpretation of two or three movement responses in an individual record, however, depends not so much on this absolute value as on the qualitative analysis of the movement answers in terms of total output, quality, originality, popularity, in terms of their projection into abnormal areas or their restriction to parts of bodies, in terms of the specific kind of content, in relation to the total number and varieties of color responses, and finally, in terms of the total personality configuration.

This is especially true of extreme deviations from the norm. Thus a score of 15 M may reflect fine imagination, stability, and integration of the personality. On the other hand, it may reflect considerable maladjustment in the form of self-preoccupation, day-dreaming, excessive fantasy, even obses-



sional or delusional developments. Again, many W's in a record may show superior abilities at abstraction and generalization. Or they may suggest extreme inner pressure, overambition and compensatory behavior. A high %F+ (Ror) may reflect fine powers of intellectual control, steadiness and clear thinking. Or it may reveal rigidity of the personality, perhaps obsessive-compulsive features or depressed conditions. It is true that these are all intuitive insights but the Rorschach is an intuitive instrument.

If we keep these very elementary facts in mind, we understand that scores falling within the normal range may not indicate normality in any area in the interpretation. In like manner, scores falling outside the range of normality may or may not reflect abnormality depending upon a qualitative appraisal.

Again, when comparing the scores made by the same individual on successive tests, the numerical differences cannot be interpreted as genuine changes in personality structure without careful qualitative analysis of the rest of the record. For example, Muench's study (114), already referred to, infers change in personality structure from a 95 per cent F+ (Ror) in the pre-test to 86 per cent in the end-test, a change which means to him lack of improvement. Again, he fails to find improvement in the case where the patient gives 36 per cent in the pre-test and 43 per cent in the end-test. This is oversimplification with all its shortcomings.

In summary, norms must be utilized in the interpretation of a Rorschach record. They should be expressed in terms of range and variability of the group. It is impossible to proceed without some standards for the group with which the individual may be compared. It must be remembered, however, that norms are only rough standards and should be used only as guides. They give important information on the central tendencies of the group. When the individual record is interpreted, however, the examiner cannot restrict himself to tables of norms. The same numerical score does not mean the same thing in every record. Scores which are numerically equal are not psychologically equal. Interpretation of all scores must be made dynamically in terms of general configurations and not absolute values. Properly used and interpreted, norms furnish the interpreter with a frame of reference for the study of the individual record.

In addition to evaluation of Rorschach scores in reference to normative material and in terms of the dynamic interrelationships of patterns in the individual record, the Rorschach interpretation involves a third phase, the study of the conclusions based on Rorschach data in conjunction with case history and other test data. The information gleaned from the Rorschach analysis is projected against the family background, education, health, his-

tory, training, social relationships, and general history, past and present. Rorschach data are re-evaluated in these terms.

The interpretation of the Rorschach record depends upon what we know about the individual. The more we know, the better we can interpret and weigh the various combinations of factors appearing in the Rorschach record and the better the over-all interpretation.

This aspect of the interpretation has been emphasized and re-emphasized by Rorschach workers in the field. Yet we are challenged again and again because we rely for our findings exclusively on the reactions to ten ink blots. Thurstone (110), for example, battles a straw man when he suggests that we claim that we can understand the past history and the dynamic personality characteristics of the individual and can predict reactivity to life situations from the responses to ink blots. No such claim has ever been made.

This brings us to a related question, what should be included in a Rorschach interpretation? From the data at his disposal, the examiner gives as detailed a study as he can of the dynamics of the personality. He tries to reconstruct the personality, to understand the intellectual functioning of the subject, to probe his inner world, to evaluate his emotions in their dynamic or disintegrative activity, and to analyze the mechanisms with which he selects and organizes his life experiences in his efforts toward self-adjustment.

The examiner may detect deviations in this functioning of his subject. He may observe certain patterns which he recognizes as characteristic of certain disorders. He identifies them. If he can, he makes recommendations as to the therapy. If possible, he tries to suggest the extent of the improvement to be anticipated. This integrated picture is based on Rorschach data and all other information available.

Two problems present themselves in this connection, that of blind diagnosis and that of differential diagnosis. Blind diagnosis based exclusively on the raw Rorschach data has no place in practical Rorschach work. This has been emphasized many times but there are still many workers who attempt to use the method in this way. Many psychiatrists demand blind analyses from the psychologist. This is unfortunate practice.

Again, it is not the primary function of the Rorschach examiner to make diagnoses in terms of the traditional clinical entities. The trend today is to question the validity of these groupings, to place less stock in the old diagnostic nosology. Unfortunately, many Rorschach examiners are diagnostically oriented. They emphasize, at times, even restrict their interpretations to nosological statements. In fact, they consider their task well done if they diagnose and classify their patients with the same label as that of the psychiatrist. This, too, is unfortunate. The Rorschach can be of much greater



value in other ways. Ross (138) has recently emphasized this point of view.

It must be recognized from the foregoing that the Rorschach examiner plays an important role in Rorschach testing. Indeed, the validity of the Rorschach results depends to a great measure on the ability, the clinical judgment, the competence, and the stability of the Rorschach examiner. Its efficient use requires training, not only in the Rorschach method, but in theoretical, experimental and clinical psychology, in psychopathology and in personality theory.

The Rorschach examiner must be objective. He must be competent. He must have clinical understanding of the Rorschach scores and of the processes revealed by them. He must be able and willing to utilize adequate norms as his guide. He must keep abreast of research and review carefully all evidences of validity of the hypotheses he uses. He must know what is theoretical, empirical, speculative, and proven.

The role of the examiner is just beginning to receive appropriate attention in another connection, in terms of the dynamics of the subject-examiner relationship. It has been pointed out that the examiner projects his feelings and his bias in the administration, in the inquiry, and in the suggestions he makes during the whole test situation. Further, he projects his feelings and his special bias in the interpretations. In this area there is need for carefully controlled systematic studies on subjective bias. Joel (80) has recently discussed this interpersonal equation in projective methods and has made interesting suggestions as to research. This is an unexplored field.

There is also great need for research on what is meant by clinical judgment and clinical intuition. As Cofer (29) has indicated, clinical judgments are based on specific factors in the testing situation, factors which can be identified and subjected to systematic study.

There is good reason why there has been increased recognition of the need for trained and qualified examiners throughout the world. More and more colleges and institutions of learning are providing opportunity for training and supervised practice.

### *Reliability*

The reliability of various aspects of the Rorschach is questioned by many psychologists. Various orthodox procedures have been utilized in the attempt to establish the reliability of the method, but have been found wanting. The adequacy and correctness of these procedures have been fully discussed by Cronbach (33). While parallel series of cards of Behn-Eschenburg

and more recently by Harrower-Erickson and Steiner (58) are now available, systematic studies have not been made to establish the fact that they really are parallel to the Rorschach blots. Indeed, the Behn-Eschenburg cards appear to be used at times as an additional test (Zulliger [193]).

The split-half method and the method of repeating the test have been for the most part discarded because they are inapplicable to Rorschach data and generally unsuccessful (Hertz [60, 70, 73], Cronbach [33]). Because of the global nature of the test, it is not possible to split it and work with isolated variables. Again, since conditions change from time to time, personality data cannot be reproduced exactly from one time to another as is anticipated when the method is repeated.

Swift (165) utilized a modification of repeating the test, studying the reliability of Rorschach categories using four different methods: (1) test-retest over a thirty-day interval; (2) test-retest over a fourteen-day interval with interpolation on the seventh day of a parallel series of blots; (3) test-retest on a parallel series after a seven-day interval; and (4) test-retest after a ten-month interval. While results varied with the methods used, she could offer data to show the reliability of the Rorschach as a clinical instrument.

The only successful approach to date to determine reliability is the method of matching which keeps the total Rorschach picture intact. Krugman (90), for example, demonstrated the reliability of the scoring and the interpretation of Rorschach records in a study of twenty problem children in which comparisons of interpretations were made by experienced judges and the response records and the scoring tabulations were matched with the interpretations. It is generally recognized, however, that this method has its limitations in that it can be applied only to small numbers and depends upon the skill of the judges used.

In this connection, studies utilizing test-retest procedures in convulsive therapy, hypnotic changes, and the like should be mentioned. Such studies, where the Rorschach test is repeated with conditions experimentally varied, indirectly prove the reliability of the method. Thus Kelley, Margulies, and Barrera (87) report that Rorschach records taken in a single day after initial electric shock (where the patient is amnesic to the test and hence cannot remember previous answers) are substantially the same. Systematic studies to determine reliability have not been developed, however, with this approach.

The problem of method to demonstrate the reliability of the Rorschach is real and challenging. No adequate statistical procedure has been suggested as yet to handle this problem. Nevertheless, it is generally felt that Rorschach interpretations possess a high degree of objectivity and reliability in the hands of skilled and experienced clinicians.



## *Validity*

Another problem always with the Rorschach examiner pertains to the validity of the Rorschach method. As has been indicated (60, 70, 73), some studies attack the problem of validity directly and attempt to establish the validity of specific patterns or specific Rorschach premises. Others, utilizing the Rorschach method for other purposes, reflect upon its validity indirectly. Still others attack the problem of basic Rorschach assumptions, without however, utilizing the Rorschach blots themselves.

In the past, validity of many of the Rorschach premises has been established largely in terms of clinical studies. It is gratifying to note that an increasing number of research workers are turning their attention to direct experimentation of Rorschach hypotheses. In some instances, experimental situations have been cleverly devised to validate various aspects of the method.

Ruesch and Finesinger (142) led the way in an early study of the relation of the Rorschach color response to the use of color in drawings, attempting an experimental verification of the affective value of color in the Rorschach. More recently, Williams (179) has shown the possibilities in this approach in his experimental study demonstrating a high significant relationship between intellectual control as indicated by the form level in the Rorschach and intellectual performance under the stress of social pressure. For another factor, form-color, the relationship was low but in the expected direction. Baker and Harris (9) report similar results in their study of the recorded speech of fourteen college students under normal and stress conditions.

Validation of some Rorschach results against laboratory procedure is also reported by Brower (17) who studied the relation of specific Rorschach factors in conjunction with other test data and cardiovascular activity before and after visuomotor conflict. He could show that some of the relationships observed tended to disappear under conflict conditions.

The nature of color and color shock has been subjected to study, utilizing the galvanic skin response technique, with conflicting results. Milner and Moreault (109) show agreement between Rorschach data and galvanometric indicators. Experimental studies by Rockwell, Welch, Kubis and Fisichelli (132) also present evidence favoring the hypothesis that inhibition of associations during color shock is accompanied by an inhibition of autonomic responsiveness as measured by the changes in the palmar skin resistance. In a subsequent study (133), the effect of color upon the associative

and autonomic responsiveness of normal persons is experimentally demonstrated.

✓ Wallen (173) using the group Rorschach, studied the nature of color shock in terms of the affective reactions of stable and unstable men in service. He concluded that color produced shock, but the shock was due not to the color *per se* but to the effect of color on the perceptual process. Color increases the difficulty of perceptual integration, induces feelings of failure on the part of insecure persons, and hence causes shock.

Lazarus (96) using slides of color and non-color series with high school students could not validate the assumption that color influences performance on the Rorschach or that shock is induced by the color on the slides. He hypothesized that shock depends not on color, but rather on difficulty, shading, and/or disturbing associations.

In like manner, the validity of the relative responsiveness to Cards VIII through X as a function of color, was not established by Sappenfield and Baker (144), who used the group Rorschach. Yet, Siipola (154) in a carefully designed experiment comparing the first conceptual reactions of normal subjects to matched chromatic and achromatic blots, confirmed the fact that color in blots induces such affective phenomena as associative blocking, strong emotional reactions, and symptoms of conceptual and behavioral disorganization. She offers the hypothesis that these symptoms of affective involvement may be viewed as an indirect effect of the presence of color, that the introduction of hue creates a situation of "hue-incongruity" which initiates conceptual conflict and emotionally-toned behavior.

Reference has already been made to Stein's experimental study (160) of personality through tachistoscopic exposure of Rorschach cards, which enabled him to make interesting observations not only on the understanding of personality but on the process of perception in the Rorschach situation.

Studies with the Levy movement cards (143) utilizing finger paintings with vague human figures are beginning to throw light on the meaning of the human movement response. Rust (143), using the Zubin M scales for the Levy blots, studied some correlates of the movement response in normal, schizophrenic and neurotic subjects and in patients with frontal ablations.

Experimental evidence of the validity of the Rorschach method is likewise furnished by the studies in which the test is given under experimentally altered conditions, demonstrating the extreme sensitivity of the method to changing conditions and to changing emotional states. Thus Stainbrook (157) could demonstrate progressive changes in Rorschach results on the basis of observations made at five-minute intervals following the onset of electric-shock convulsions. Again, Morris (112) reported reliable changes in pre- and post-treatment records in patients subjected to metrazol therapy.



The sensitivity of the method to mood changes and to suggestions induced under hypnosis was demonstrated by Levine, Grassi, and Gerson (97), who used the verbal graphic Rorschach methods. Bergman, Graham, and Leavitt (14) studied the Rorschach reactions in consecutive hypnotic age level regressions. Lane (94) validated interpretations of the Rorschach movement factor by inducing creativity and introversive mechanisms by hypnotic suggestion in a non-productive subject. Personality changes after the administration of drugs such as histamine were also reported by Robb, Kovitz and Rapaport (131). Wilkins and Adams (178) likewise demonstrated personality changes in patients under hypnosis and sodium amytal, on the basis of the Rorschach record.

Such experimental studies as suggested above mark only a beginning in establishing the validity of the Rorschach method. They show great promise, however. It is hoped that other aspects of the Rorschach will be subjected to this kind of experimental approach so that many of the hypotheses implied in Rorschach interpretations may be experimentally verified.

Most of the Rorschach studies on validity, however, are based on the comparisons of contrasted groups and on case studies, especially those using "blind diagnosis" and the matching method.

In the method of compared groups, the Rorschach has been shown to differentiate between individuals of varying age, intelligence, background, school achievement, of different race or nationality, of deviated personality, and between individuals suffering from various kinds of mental disorders. Many of these studies consist of analysis of the scores of extreme or contrasting groups. Others use the method of equating groups for various factors and identifying batteries of Rorschach factors which appear significantly more frequently in one group than in another. Many of these studies have been reviewed in the literature (Hertz [60, 70], Hertz and Ellis [73], and Bell [11]).

Some of the more important studies utilizing this approach include those on non-reading and clinic children by Vorhaus (172), adjusted and mal-adjusted children by Davidson (36), stutterers and non-stutterers by Krugman (91) and Melzer (107), and institution and foster-home adolescents by Goldfarb (48).

Important studies have been offered by Werner (175) on brain-injured and non-brain-injured mental defectives, by Sarason and Sarason on groups of high-grade familial defectives (145) and cerebral palsied defective children (146). Sloan (155) compared two groups, one who had been legally committed as subnormal but who did not appear to be psychometrically defective, and a second group with high-grade or borderline intelligence, psychometrically. He demonstrated that affective factors may interfere with the

full utilization of capacities. Jolles (81, 82) likewise emphasized emotional and personality maladjustment as an important component in mental deficiency on the basis of a Rorschach study of sixty-six children who had low ratings in psychometric tests. The method of group comparison was also used by Abel (1), who studied the relationship between academic success and personality organization, by matching fifteen pairs of moron girls on the basis of IQ and chronological age, where there was a difference in school placement between the members of each pair.

Most recently, there has been an extension of this technique in the development of "signs" which occur more frequently in one group than in a controlled or contrasting group. Thus signs for evaluating good adjustment in school children have been developed by Davidson (36, 37), signs indicating mental deficiency by Sloan (155) and Jolles (81, 82), signs of adjustment in adults by Muench (114), signs to determine neurotic involvement by Miale and Harrower-Erickson (108), signs of the schizophrenic process by Klopfer and Kelley (89), signs of organic impairment by Piotrowski (118), Aita, Reitan and Ruth (2), Armitage (8), and Hughes (77). Signs have even been developed to discriminate between outstanding and non-outstanding mechanical workers by Piotrowski, Candee, Balinsky and Holtzberg (125).

Munroe's check list (115) is really a series of signs providing for the notation of deviations which are based on clinical experience. Munroe claims that the check list facilitates rapid inspection of significant aspects and interrelationships on which a general clinical judgment may be used. Using the inspection technique, she could differentiate between various degrees of adjustment of her college students.

Ross and Ross (140) developed a general "instability" and "disability" rating, consisting of combined and weighted signs occurring more often in neurotic and organic subjects than in controlled subjects, the ratings being validated against clinical findings and Binet subtests.

The signs of Buhler, Buhler and Lefever (22) which have already been referred to, probably represent the most extreme form of describing personality on a statistical basis. The authors have devised a numerical score, the Basic Rorschach Score, which evaluates personality disintegration in terms of conflict, impairment, and reality loss. According to the authors, the score differentiates the various diagnostic groups. The possibilities and limitations of this approach have been considered in a recent symposium (177) and also by Cronbach (33).

Negative results have also been reported in studies attempting to utilize signs. Kurtz (93) could find no evidence that signs are valid in the selection of personnel in the various occupations, in industrial or in military occupations. McCandless (104) failed to differentiate between two groups of officer



candidates who differed widely in academic progress and achievement, using the Munroe check list. Cronbach (35) failed to confirm Munroe's findings using a similar group of college students and failed to establish her signs as predictive of academic success.

Validation of the Rorschach is also frequently made in terms of comparison with outside criteria, case records, other test data, interviews, teachers' reports, psychoanalytic data and the like. At times these comparisons are exclusively qualitative. On the other hand, frequently an adequate number of cases permits quantitative techniques to be applied. This approach is often impressive especially when blind personality analyses are made or when the matching techniques are utilized and/or when elaborate statistical procedures are employed.

Thus validation using the results of other objective personality tests have been reported by Wishner (181) who established the validity of such Rorschach factors as R, W, and Z on the basis of their relation to the Wechsler-Bellevue score with neurotic patients. He failed, however, to validate the %F+. Burnham (23) studied the degree of relationship between the % H and the Wechsler-Bellevue Picture Arrangement scores. The MMPI scores were utilized by Altus and Thompson (5), who reported a high correlation of items in the Group Rorschach and the Schizophrenic Scale scores, by Clark (28), who could validate the Rorschach color interpretations in the Group Rorschach, and by Thompson (169), who in like manner studied the MMPI correlates of two types of Rorschach movement responses in the Group Rorschach for two college groups.

Correlation of results with independent and experimentally controlled behavioral criteria has yielded conflicting results. Thus behavioral measures obtained from a teachers' rating scale and from parent interviews were employed by Swift (166) to validate Rorschach measures of insecurity in terms of ratings and "signs" but with generally negative results. When she adopted a more global approach (164), however, and matched teachers' descriptions of the personalities of her thirty preschool children with Rorschach analyses, a significantly high number of correct matchings was made.

Hertzman and Pearce (75) utilized behavior in therapy, dream interpretations, self-descriptions and descriptions of the subjects by others, in order to validate the meaning of the human figure response in the Rorschach. The judgments of therapists were utilized by Wheeler (176) in his analysis of Rorschach indices of male homosexuality. Garfield (45) took as his criteria the clinical diagnoses of a staff and demonstrated the validity of Rorschach diagnoses. Again, psychoanalytic data were utilized liberally by Schachtel in his various studies on the symbolism of form (147), movement (150), color

(148) and situational influences (149). At no time, however, does he present systematic studies of his interesting Rorschach hypotheses.

Case studies demonstrating the clinical validity of the Rorschach abound in the literature. "The case of Gregor" (12), presented at a recent symposium of the APA, which included a report of twenty-seven different projective techniques, one of which was the Rorschach, is illustrative of this approach.

Studies employing the method of correct matching have demonstrated a high degree of validity for the Rorschach interpretation. Patterson and Magaw (117) matched the Rorschach pictures of institutionalized defective boys with personality sketches written by trained observers with a significant number of correct matchings. Krugman (90) obtained a highly satisfactory result by matching Rorschach analyses with clinical case study abstracts, matching scoring tabulations with interpretations, thereby establishing a high degree of objectivity and clinical validity for the Rorschach. Reference has already been made to Swift's success in matching Rorschach analyses with teachers' descriptions (164).

Validity relying almost exclusively upon statistically designed procedures has been reported by various statisticians and research workers, interested in an ultraquantitative approach to the Rorschach method. Some have tried to intercorrelate various Rorschach factors or correlate Rorschach indices with total score and apply techniques of item analysis. Others have used factor analysis to determine which combinations carry particular loadings. Thus, using factor analysis of data provided by the Harrower-Erickson Multiple Choice Check List, Wittenborn (184) concluded that the abstract scoring procedures usually employed are of no value in attempting to appraise the behavioral significance of Rorschach responses elicited by check list procedures. A later study (183), a factor analysis of discrete responses to the Rorschach blots, again failed to support some of the common interpretations for Rorschach factors. Other studies utilizing a variety of analytical designs, tested out other Rorschach hypotheses (182, 186). He could show that Rorschach responses differ from each other with respect to the degree of perceptual control characterizing them (183, 185). He also could present evidence for the validity of scoring the human movement response as reflecting an important feature of personality not predicted from the color responses, and of combining other color responses and interpreting them differently from the total movement score. He likewise established the validity of interpreting responses separately from different areas of the card. He could not, however, justify the practice of working with the more refined scoring categories of M, C, and details.

Again, Hughes (77) proposed twenty-two different Rorschach signs for



the detection of organic brain pathology, based on a factor analysis technique. Fourteen signs were established as statistically diagnostic for a group of 218 patients. Hsü (76) likewise attempted a factor analysis of Rorschach factors in order to study the reliability of the method.

Another approach utilized in the attempt to demonstrate the validity of the Rorschach method appears in those studies which demonstrate the power of the Rorschach method as an instrument of prediction. Thus Munroe (115) reported a high degree of success in predicting the adjustment of college students, academic failure, referrals to psychiatrist, and problem behavior observed by teachers from Rorschach data using the Rorschach Group Method and the Inspection Technique of scoring. Piotrowski (119, 120) had significant success in predicting the effectiveness of insulin treatment on the basis of the analysis of pretreatment Rorschach records. Similarly, prognostic patterns were identified by Halpern (54) for the prediction of response of schizophrenic patients to therapy, and by Morris (112) for the prediction of the outcome of treatment based on metrazol. Siegel (153) identified Rorschach factors associated with improvement and non-improvement and found them valid for prediction in a child guidance clinic, and Bradway (16), working with "promiscuous" girls, identified a battery of patterns which were of prognostic value in determining treatability.

Using the group Rorschach, Montalto (111) could predict achievement of women in college in terms of "signs" of adjustment. Shoemaker and Rohrer (152) predicted success of students in the study of medicine on the basis of differential Rorschach results for "over" and "under" achievers. On the basis of clusters of Rorschach patterns obtained from group Rorschach records, Thompson (168) demonstrated the value of the Rorschach in predicting success in college.

In 1942, Piotrowski (121) reviewed the progress which had been made in personality study with the Rorschach method, and emphasized especially some of the valid predictions which could be made of personality development. More recently, Morris (113) in a carefully designed study demonstrated the power of the Rorschach to predict personality attributes. Again, the prognostic possibilities of the method were demonstrated by Hertz (67) who identified ten configurations in terms of Rorschach patterns which were studied both quantitatively and qualitatively and which revealed suicidal tendencies. These configurations were further analyzed and verified in a follow-up cross-validation study (68).

It may be concluded from the above brief summary that various kinds of research designs have been utilized in continuing the attack on the problem of the reliability and validity of the Rorschach method. It would appear that many aspects of the Rorschach are impressively supported by a host of

validating studies of wide variety. Nevertheless, much of the research has serious limitations; many of the findings are inconclusive. In reviewing the material, one finds that several important problems suggest themselves.

### *Problems*

First, most of the studies have never been replicated. This is a serious omission. It is important to determine whether the results obtained in one study are peculiar to the sample used. We know, for example, that results often may be explained in terms of the statistical method which has been employed in analyzing the data. They may be explained also in terms of the experience, the clinical insight and ingenuity of the investigators. Thus Munroe's work (115) in predicting academic success would be much more valuable if it were verified. In a recent study, Cronbach (35) attempted to repeat her study on a similar college group. He could not duplicate her results. Again, Muench's study (114) validating Roger's non-directive method of therapy could not be repeated by Hamlin, Albee and Leland (55) or by Carr (24). We have many valuable studies which, if repeated, would contribute immeasurably to the scientific status of the method. They await replication.

Again, much is in use in the Rorschach which has never been validated, even clinically. Many interpretations which we give to Rorschach patterns have not been systematically validated in themselves. Some introduced tentatively as working hypotheses still retain their hypothetical status. These include the various kinds of rare-detail categories which appear in the Klopfer and Kelley manual (89), the *m*, the texture response, animal movement, *k*, *Ch'* (as used in Hertz scoring) or *C'* (as used by Klopfer), and the like. Several patterns which we use call for systematic verification. These include *W : M*, *P : O*, the so-called circle of refinement (89), and various other combinations. In 1943, Piotrowski (122) published a stimulating paper on tentative Rorschach formulae for educational and vocational guidance in adolescence. Most of these formulae are still tentative, despite the fact they are used as established facts by many Rorschach workers.

Again, the Rorschach test is frequently employed as an instrument of research. While several aspects of the method have been sufficiently validated to permit its use for research purposes, modified procedures and tentative findings based on small scale studies and/or on isolated case studies are uncritically applied. This is especially true of many of the studies utilizing the Group Method and the Multiple-Choice Method.

In this connection, it should be emphasized again that despite the value



of the method of comparing groups and of the matching technique, the individual Rorschach patterns have not been validated. In comparing groups, the Rorschach method may be shown to be valid as a differentiating instrument. The individual Rorschach pattern has not been shown to be valid, however. The factors or patterns or signs which appear with greater frequency in one group than in another may be said to be associated with the dominant personality patterns of that group and inferentially to be reflective of them. But from this we cannot conclude that each pattern is of itself valid. This is frequently forgotten.

This may likewise be said of the matching technique. Even though the method as a whole may be shown to be valid through correct matching, it must be emphasized that even in successful matching, each item in the description has not been demonstrated to be valid. In fact, although many items in themselves may be incorrectly matched, the total picture may present a correctly matched whole. This has been emphasized by many reviewers (Cronbach [33], Rotter [141]).

This poses a practical problem to those of us who work with the Rorschach as to what to utilize in the interpretation and what not to use. The examiner must have sufficient experience, knowledge and discernment to differentiate between empirical data and systematically verified data. More important, he must be on the alert to distinguish valid inferences from highly personal and sometimes fanciful interpretations.

We have already suggested a major problem in Rorschach research, a problem which has been discussed again and again but which is ever prevalent, that is, the use of Rorschach scores as separate and independent items with specific meanings, despite the recognized importance of dealing with them in their various interrelationships and in terms of the record as a whole. This point has been emphasized and re-emphasized but the procedure of using single factors in isolation persists. Recently, patterns of scores or composite scores have been recommended (Hertz [66], Rapaport [129]). Thus formulae as  $(FC - [CF + C])$  wt,  $(FCh - [ChF + Ch])$  wt,  $(M + Cwt)$  are used in Rorschach interpretation. Even with composite scores, however, it is difficult to get meaningful results when they are used in isolation, because composite scores also depend upon other scores and the record as a whole. Hence studies which isolate composite patterns for study are not generally successful.

More recently, a configurational approach is recommended by Hertz (67) which relies upon 1) normative data for Rorschach factors and patterns which are evaluated, however, in terms of the individual record, and 2) clinical interpretation and judgment. The configurational approach combines the qualitative and quantitative evaluation and is the procedure

which is usually used when interpreting a Rorschach record. The position is taken that if the configurations are carefully and explicitly described, they may be used by other clinicians with reliability, objectivity and validity, and may be subjected to experimental investigation. Thus Hertz (67) demonstrates that suicidal trends may be detected in Rorschach records on the basis of the evaluation of certain configurations which are explicitly described. She emphasizes, however, that this subjective-objective procedure must be subjected to further research to determine whether other test-interpreters will be able to make the same predictions.

Of all the methods of validation mentioned, probably the most promising is the method of prediction. Unfortunately, few of the studies using this method have been repeated. As already indicated, studies in which attempts have been made to repeat the procedure have met with little success. If this method can be systematically developed, however, so that Rorschach test interpreters with sufficient training and experience can make the same predictions on the same data, the Rorschach method will be on firm and valid ground. Of course, in this approach, it is important also that the specific patterns and their differentiating criteria, and if possible, the bases for clinical judgments be explicitly described and where possible, experimentally established.

There is an additional problem which has caused many of us much concern—that is, the validity of the procedure and the studies designed to streamline the Rorschach method. Of late there has been much concentration on shorter procedures, mass methods, rapid inspection scoring and diagnostic sign batteries in order to permit mass testing, mass diagnosis, mass production of Rorschach interpretations and even mass research.

✓ In order to cover more subjects in faster time, the Rorschach Group Method has been used extensively despite the fact it has not been developed beyond its early experimental status. Several variations of the original method of Harrower-Erickson and Steiner (57) are used but to date no group method has been reliably established as valid; scoring norms for such factors as normal details, popular responses, form quality, and the like have not been determined; and adequate group norms have not been amassed for the age groups which are studied by means of the Group Method. It has not even been systematically established that the same principles of interpretation upon which we proceed with the individual record operate with the group record. Despite these limitations, which have been emphasized again and again in the literature (Challman [27], Hertz [64]), the Group Method has been applied extensively in clinical, educational, and industrial studies, in the armed services, and even as an instrument of research. The reviews of Munroe (116) on the use of the Rorschach in college counseling, of Har-



rower-Erickson (56) on the general application of the group method, of Piotrowski (123) on the use of the method in vocational selection, all emphasize the extensive use which has been made of the Group Method. More significantly, it is used extensively as an instrument of research. Roe (134-137) for example uses it as one of her techniques in her important research project on personality and vocation.

Results in studies using the Group Method are conflicting. Abt (2) reports data showing the group Rorschach as an efficient instrument for psychiatric screening of Marine Corps recruits. Harrower and Cox (162) apply the group method to various occupational groupings and find it highly valuable for selecting and placing the worker in industry. Steiner (161-162) reviews the literature and concludes that the group method is valuable in differentiating between successful and unsuccessful workers, in understanding the personality dynamics underlying occupational adjustment, and in differentiating general personality patterns for various occupational groups.

Again, Stainbrook and Siegel (158) found it valuable in making a comparative study of southern Negro and white high school and college students. Thompson (168) and Montalto (111) could differentiate between achieving and non-achieving college students by means of the Group Method. The Group Method was also utilized with apparent success by Kabach (83) who studied the relationship between personality and vocational choice, and by Brozek, Guetzkow, and Keyes (18) in their study of changes in nutritional status and personality. Anderson and Munroe (6) utilized the group procedures to study personality factors involved in student concentration on creative painting and commercial art. Reference has already been made to the extensive studies by Roe in which she utilized the Group Method among other techniques for the study of the personalities of artists (135), scientists and technicians (134) and biologists (137).

On the other hand, generally negative results are stressed by other investigators. Kurtz (93) finds the Group Method without validity in the application to industrial and personal problems. In most of the studies which he reviews, he fails to find any justification for acceptance of the group Rorschach as a valid instrument. Anderson (7) failed to confirm the study by Piotrowski et al. (125) and could not predict the efficiency ratings of machinists on the basis of group Rorschach records. Ross, Ferguson and Chalke (139) report that the group Rorschach has limited use as an aid in officer selection in the Canadian army, despite the positive results which were reported by Harrower-Erickson (56). In a more recent study, Cronbach (35) could not establish the validity of the group Rorschach in his attempt to correlate Rorschach adjustment scores with the ratings of emotional adjustment by heads of dormitories and with sociometric ratings.

In the light of the few attempts to subject the Group Method to systematic research and to establish normative data, it must be concluded that the application of the method has been indiscriminate and unwise. The studies which report positive results should be repeated, especially those which claim to be able to screen effectively and to reflect differentiating personality patterns for various vocations and arts. It is hoped that both the positive and negative results will stimulate further refinements and modifications in the group procedure and conscientious research for making it a reliable and valid instrument.

Less promising results have been reported for the Multiple Choice Test as a screening device. Harrower-Erickson (56) has used the technique with most satisfactory results. More recently, Cox (30) reports that successful sales clerks could be selected on the basis of this technique, and has devised a scoring key made up of the items which differentiated the top from the bottom quarter of a group of 108 sales clerks.

On the other hand, Challman (26) had little success with this method as a screening device. Wittson, Hunt and Older (187) concluded on the basis of a study of three groups of naval men that it is unsuitable for military selection. Jensen and Rotter (79) were unsuccessful in screening officer candidates. According to Wenfield (180) the test did not screen maladjusted from adjusted women in military service. Similarly, Springer (156) found it unsuccessful in screening naval personnel.

Again, Malamud and Malamud (101) could not use the Multiple Choice Method to discriminate between normal subjects and psychiatric cases. Engle (40) also reported that it was inadequate for differentiating between well-adjusted and maladjusted high school pupils.

In summary, it must be recognized that the validity of both the Rorschach Group Method and the Multiple Choice Test is far from established. Despite the positive results reported in some studies, both techniques require considerable experimental work in order to place them within the bounds of reliability and validity. It is possible that with further refinement and modification they will attain greater validity. One modification of the method, the Rorschach Ranking Test, has been developed by Eysenck (41, 42) who reports that this technique along with three other tests can reliably discriminate neurotic and normal subjects. Suggestions for modifications and revisions have also been made by Malamud and Malamud (102), Challman (27), and Lawshe and Forster (95). Kellman (85) is in the process of an elaborate revision of the Multiple Choice. It should be emphasized, however, that until such developments take place and objective evidence of the validity of the procedures is presented, they must be used with caution.



They should not be employed as instruments of research in their present state.

Another attempt to objectify Rorschach data and speed up the process of interpretation is the sign approach already referred to. Studies identifying signs of adjustment, neurotic trends, organic impairment, schizophrenia, and the like have been developed with more or less acceptable statistical validity. Indeed, we are told that the sign approach is simple and well adapted to the Rorschach (Cronbach [33]). Statisticians apparently like it because it represents the statistical approach to the Rorschach par excellence. While this approach may be simple, in our judgment it is not only unadapted to the Rorschach but it is incompatible with the basic principles of the method.

Clinically, the battery of signs is not valid since it fails to take into account the dynamic relationships among the various psychological processes for which the so-called signs stand. Just as with other scores, the signs depend for their real significance on other patterns and on the total record. Even for superficial evaluations, personality configurations must be studied and not independent signs.

A very serious limitation to most batteries of signs is the exclusion of patterns which by their presence suggest pathology. For example, contaminations, variations in form level, in originality and in output, position responses, color deterioration, and the like are included in few lists of signs, with the exception of the schizophrenic battery. It has been demonstrated again and again that schizophrenic records may show the necessary number of signs pointing to adjustment, for example, and yet have one or more of the above indicators which should at once withdraw the label of adjustment (55, 186).

Again, in many batteries, signs are added to obtain a composite score. Again we have the atomistic approach, do we not? We have, in fact, an appraisal of the whole in terms of the summation of individual parts. The fact that the parts are called "signs" does not change the case. Our subject has again become a collection of discrete test scores and we have again lost sight of the larger framework of relationships.

Despite the acceptance of the sign approach, study after study illustrates the difficulties of treating scores separately. Siegel (153), who developed a battery of signs reflecting adjustment in order to predict response to continued psychotherapy, cautions that the signs must be used judiciously within the essential configuration of the test pattern. Rabin (127) emphasizes the inadequacy of the sign approach for differential diagnosis. Piotrowski (124) who pioneered in the development of signs, prefers a more systematic method which he called "perceptanalysis" which emphasizes the dynamic

interrelationships of the underlying forces in the personality and which is based on the %F+ as a point of reference. Examination of his method shows it to be a configurational approach and not a sign approach at all.

Munroe's check list (115) has not been widely used or successful in application. It is difficult to use even for trained examiners. In our experience in the process of rapid evaluation, many of the items are judged wrong. In a rapid over-all inspection such as is suggested, certain items serve as the basis for the decisions. Even a few errors then make considerable difference in the results.

It may be that those who are successful with the inspection technique may attribute their success to experience, insight and intuition rather than to the checks or signs per se. The inexperienced clinician cannot use the check list with any degree of validity.

The problem in the use of signs is a serious one. The sign approach is designed to permit the inexperienced worker to use—or should I say, misuse the Rorschach method. Munroe tells us:

The adjustment rating . . . is largely independent of the special skill and special experience of the individual examiner. It can be made quickly from spontaneously produced Rorschach responses according to clinical criteria familiar to examiners with even moderate Rorschach training. There is indeed some evidence to suggest that a quantitative score may be safely substituted for the examiner's rating, thus reducing the factor of personal judgment to a marked degree (115, p. 11).

Again, Davidson is happy to relate

. . . fifteen of the seventeen [signs] may be tallied by a clerk while only two (color shock and shading shock) require the judgment of an experienced Rorschach examiner (37, p. 33).

The various studies based on short cuts and sign approaches cannot be considered clinically valid or acceptable. With such approaches as described, patterns are applied mechanically and all evaluation of the dynamics of personality is excluded. This is a distortion of the Rorschach method. Clinically, results are sterile. Such attempts at oversimplification are understandable in terms of a desire to give wider utility to the method. Despite temptation, however, it is important to remember that the rule about royal roads applies to Rorschach as well as to other disciplines of learning. Those who emasculate the method with the view to giving it to clerks to handle are doing much to keep the Rorschach from attaining full scientific status.

Finally, one more problem should be considered, the extent to which statistical procedures should be applied to various kinds of Rorschach data. As already indicated, various techniques have been utilized with more or less success.



It has been amply emphasized in the literature that because of the nature of the Rorschach, most of the orthodox statistical procedures which have been used to date with Rorschach data are highly inappropriate (Cronbach [31-34], Hertz [60, 70], Frank [44], Krugman [92], and others). Many of the traditional procedures have been misapplied, often resulting in erroneous conclusions. In other cases negative results might have appeared more favorable if appropriate methods had been used. Cronbach (33) is well within the bounds of accuracy in pointing to the errors that have crept into various efforts in the field which have sought to apply orthodox statistical methods to Rorschach data.

Throughout the years, these techniques have been the subject of much concern to many of us. Investigators have shown that many of the Rorschach factors do not follow the normal distribution and hence procedures based on the assumption of normality are fallacious when applied to them. With many factors, like C, CF, m, c, cF, Do, S, s, and others, so few appear that frequently from 75 to 95 per cent of the subjects fail to score in these items. Yet means and medians for these factors are computed again and again in research.

There can be little doubt that where appropriate, the application of statistical procedures and statistical controls adds to the objectivity and validity of the method. Yet, even where appropriate statistics have been applied and results shown to be highly reliable, the interpretation of the results has often been fallacious and their application in serious error. Too often, there has been so much concentration on numbers and their statistical manipulation that meaningfulness of the data has been forgotten. While no criticism or disrespect is intended for many carefully planned statistical studies, we sometimes have the feeling that investigators have become so absorbed in their calculations and statistical formulae that important developments with the Rorschach have been impeded rather than encouraged.

Again, frequently statistical reliability gives the investigators a feeling of security, so much so that they proceed to make clinical inferences and deductions without consideration of the global nature of the instrument. No matter how proper or correct the statistical procedure, results are of little value unless they are transformed into configurational and dynamic interpretations.

In some instances, statistically reliable results based on one kind of sample are applied to other groups and to other individuals, indiscriminately. It is elementary to mention that the results of a study depend upon the nature of the sample used, age, education, kind of background, vocational experience and a host of other factors. Yet so many Rorschach examiners have applied results of published studies without investigating the

nature of the sample on which they are based, or if they do investigate, they are not deterred in applying Rorschach results to their sample regardless of fit or propriety. Ford (43) develops Rorschach norms for children, yet uses the scoring criteria which are based on adolescent and adult populations. Harrower-Erickson and Miale (108) present norms based on the group Rorschach, yet express them in terms of scoring criteria developed for the individual Rorschach. Rapaport et al. (129) present elaborate statistics showing characteristic Rorschach reactions for groups of neurotic and psychotic patients, comparing them with a "normal" group, this latter composed entirely of a small group of constricted highway patrolmen. Hallowell (51) who has employed the Rorschach extensively to investigate personality and culture of non-literate groups is of the opinion that valuable insights are possible even in cases where *there are no group norms*. Indeed, in many studies on the acculturation process (53), norms and "signs" of adjustment are applied to different cultural groups, despite the fact they were developed on the basis of records of children and adults in this country.

It does not make this procedure any less reprehensible when the investigator recognizes and admits the impropriety of his procedure. The fact remains, he proceeds to use it and makes certain inferences which are not justified. It is because of this improper application of Rorschach results that so many records have been misinterpreted and so many results erroneously reported.

### *Summary*

In summary, Rorschach workers are increasingly aware of and concerned with problems of theory, technique and methodology, problems of objectification and verification, and problems of application. Various kinds of research designs have been utilized in the attempt to meet these problems. Some progress has been made in systematizing research procedures, amassing scoring criteria and norms, using more scientific methods in handling data, examining results more critically and developing new methods for handling the complex Rorschach material.

As a result, more and more aspects of the Rorschach have been validated. Probably the studies on validation which show the greatest promise include the experimental investigations of basic Rorschach postulates and those designed to demonstrate the power of the method to predict personality development and reactions to various kinds of therapy. Clinical validation must not be discounted, however. Carefully designed clinical research has led



to the establishment of the validity of many of the Rorschach findings which are accepted today.

On the other hand, there has been a serious lag in basic research. Many of our hypotheses have not been clarified or validated; many of our studies are inadequate. Studies which show promise have not been replicated. In many studies, statistical procedures have been erroneously applied. In others, statistical procedures have been correctly applied but they have been over-emphasized and/or the results have been misinterpreted. Finally, too much attention has been given to the exposition of short cuts and to their application despite the lack of evidence of reliability and validity.

There can be little doubt that there is still much to be done by way of objectification and verification of the Rorschach method. Most encouraging, there is no lack of will to do it.

It is obvious that there must be a complete re-evaluation of the research design and of the statistical procedures appropriate for the treatment of Rorschach data. Indeed, the problem goes beyond the Rorschach method. It is the problem of all techniques which attempt to describe and evaluate and "diagnose" personality. It is a problem which has been considered again and again in studies, discussions and symposia (Zubin [188-190]). It is generally emphasized that new methods must be explored which may be applied to the Rorschach method and to all other personality methods without sacrificing values of qualitative analysis. New methods must be forged which will keep the individual intact when the various aspects of his personality are subjected to scrutiny.

Allport (4) early emphasized the need for the ideographic approach to the study of personality and the systematic development of the individual case study. Perhaps this should be explored further. Other possibilities which have been suggested for use in conjunction with the Rorschach include the further development of factor analysis, the analysis of variance and covariance, and Fisher's method of discriminant functions (Zubin [190]). The Q technique as described by Stephenson (163) may likewise be applied to Rorschach data. Another approach may be in the further development of Zubin's early rating scales for classifying Rorschach categories (191). Cronbach, too, has suggested two other statistical procedures. His method for pattern tabulation (31, 32, 34) may be of interest, although only two or three scores can be handled at a time. He also has developed a variation of the blind matching method which, though highly involved and laborious, has shown some interesting possibilities for the validation of qualitative analyses of personality structure. All of these newer statistical methods are being explored today in conjunction with Rorschach research. Statisticians are joining with Rorschach workers in this important task of developing new

procedures which will permit quantitative analysis of the personality structure with appropriate cognizance of the uniqueness of the individual personality.

The problems we have discussed offer challenging opportunities for further work. Despite these problems, however, research on the Rorschach has made tremendous strides. It is fair to say that research to date provides clinical, experimental, and statistical evidence of sufficient importance to justify favorable regard for the method as a clinical instrument. Despite our limitations in theoretical explanation and in statistical verification, those of us in clinical work know that we have an instrument which works under the critical eye of the clinician. I think it fair to say that the only time it does not work is when it is dissected, distorted, modified, objectified to the point of sterility, and subjected to piecemeal and rigid statistical manipulation. Otherwise it works. The task for the Rorschach worker, for the statistician, indeed, for all who are interested in personality theory and projective methods is to find out why.

## REFERENCES

1. ABEL, T. M. The relationship between academic success and personality organization among subnormal girls. *Amer. J. Ment. Def.*, 1945, 50, 251-256.
2. ABT, LAWRENCE EDWIN. The efficiency of the Group Rorschach Test in the psychiatric screening of Marine Corps recruits. *J. Psychol.*, 1947, 23, 205-217.
3. AITA, JOHN A., REITAN, RALPH M., AND RUTH, JANE M. Rorschach's test as a diagnostic aid in brain injury. *Amer. J. Psychiat.*, 1947, 103, 770-779.
4. ALLPORT, GORDON W. *Personality*. New York: Henry Holt and Company, 1939.
5. ALTUS, W. D., AND THOMPSON, GRACE M., The Rorschach as a measure of intelligence. *J. Consult. Psychol.*, 1949, 13, 341-347.
6. ANDERSON, IRMGARD, AND MUNROE, RUTH. Personality factors involved in student concentration on creative painting and commercial art. *Rorschach Res. Exch.*, 1948, 12, 141-154.
7. ANDERSON, ROSE G. Rorschach test results and efficiency ratings of machinists. *Personnel Psychol.*, 1949, 2, 513-524.
8. ARMITAGE, STEWART G. *An Analysis of certain psychological tests used for the evaluation of brain injury*. Psychological Monographs, Vol. 60, No. 1. Whole No. 277, Washington, D.C.: Amer. Psychological Association, Inc., 1946.
9. BAKER, LAWRENCE M., AND HARRIS, JANE S. The validation of Rorschach Test results against laboratory behavior. *J. Clin. Psychol.*, 1949, 5, 161-164.
10. BECK, S. J. *Rorschach's Test*. New York: Grune & Stratton, 1949. Vol. I—Basic Processes. Vol. II—A Variety of Personality Pictures.



11. BELL, JOHN ELDERKIN. *Projective Techniques*. New York: Longman's, Green & Co., 1948.
12. BELL, JOHN ELDERKIN. The case of Gregor: psychological test data. *Rorschach Res. Exch.*, 1949, 13, 155-205.
13. BERGMANN, MARTIN S. Homosexuality on the Rorschach Test. *Bull. of the Menninger Clinic*, 1945, 9, 78-83.
14. BERGMANN, M. S., GRAHAM, H., AND LEAVITT, H. C. Rorschach explanation of consecutive hypnotic age level regressions. *Psychosom. Med.*, 1947, 9, 20-28.
15. BOCHNER, R., AND HALPERN, F. *The Clinical Application of the Rorschach Test*. New York: Grune & Stratton, Inc. 1945, 2nd ed.
16. BRADWAY, K. P., LION, E. G., AND CORRIGAN, H. The use of the Rorschach in a psychiatric study of promiscuous girls. *Rorschach Res. Exch.*, 1946, 10, 105-110.
17. BROWER, DANIEL. The relation between certain Rorschach factors and cardiovascular activity before and after visuo-motor conflict. *J. General Psychol.*, 1947, 37, 93-95.
18. BROZEK, J., GUETZKOW, H., KEYS, A., CATTELL, R. B., HARROWER, M. R., AND HATHAWAY, S. R. A study of personality of normal young men maintained on restricted intakes of vitamins of the B complex. *Psychosom. Med.*, 1946, 8, 98-109.
19. BRUNER, J. S., AND GOODMAN, C. C. Value and need as organizing factors in perception. *J. Abnorm. Soc. Psychol.*, 1947, 42, 33-44.
20. BRUNER, J. S., AND POSTMAN, L. Tension and tension release as organizing factors in perception. *J. Personality*, 1947, 15, 300-308.
21. BRUNER, J. S., AND POSTMAN, L. Symbolic values as an organizing factor in perception. *J. Soc. Psychol.*, 1948, 27, 203-208.
22. BUHLER, CHARLOTTE, BUHLER, KARL, AND LEFEVER, WELTY D. *Development of the basic Rorschach score with manual of directions*. Los Angeles, Calif.: Rorschach Standardization Studies, No. 1, 1948, ix, 190 pp.
23. BURNHAM, CATHERINE A. A Study of the degree of relationship between Rorschach H% and Wechsler-Bellevue picture arrangement scores. *Rorschach Res. Exch.*, 1949, 13, 206-209.
24. CARR, ARTHUR C. An Evaluation of nine non-directive psychotherapy cases by means of the Rorschach. *J. Consult. Psychol.*, 1949, 13, 196-205.
25. CATTELL, R. B. *Description and Measurement of Personality*. New York: World Book Co., 1946.
26. CHALLMAN, R. C. The Validity of the Harrower-Erickson Multiple Choice Test as a screening device. *J. Psychol.*, 1945, 20, 41-48.
27. CHALLMAN, ROBERT C. Book Review: Large scale Rorschach Techniques. A manual for the Group Rorschach and Multiple Test, by Harrower-Erickson, M. R., and Steiner, M. E. *Psych. Bull.*, 1946, 43, 285-287.
28. CLARK, JERRY H. Some MMPI correlates of color responses in the group Rorschach. *J. Consult. Psychol.*, 1948, 12, 384-386.
29. COFER, C. N. An analysis of the concept of "clinical intuition." In Kelly, G. A., *New Methods in Applied Psychology*, 1947. Pp. 219-227.

30. COX, KENNETH J. Can the Rorschach pick sales clerks? *Personnel Psychol.*, 1948, 1, 357-363.
31. CRONBACH, LEE J. A validation design for qualitative studies of personality. *J. Consult. Psychol.*, 1948, 12, 365-374.
32. CRONBACH, LEE J. "Pattern tabulation": a statistical method for analysis of limited patterns of scores, with particular reference to the Rorschach test. *Educ. Psychol. Measmt.*, 1949, 9, 149-171.
33. CRONBACH, LEE J. Statistical methods applied to Rorschach scores: a review. *Psychol. Bull.*, 1949, 46, 393-429.
34. CRONBACH, LEE J. Statistical problems in Multiscore Tests. *J. Clin. Psychol.*, 1950, 7, 21-25.
35. CRONBACH, LEE J. Studies of the Group Rorschach in relation to success in the college of the Univ. of Chicago. *J. Educ. Psychol.*, 1950, 41, 65-82.
36. DAVIDSON, H. H. *Personality and economic background: a study of highly intelligent children*. New York: Kings Crown Press, 1943.
37. DAVIDSON, H. A measure of adjustment obtained from the Rorschach protocol. *J. Projective Techniques*, 1950, 14, 31-38.
38. DUE, FLOYD O., AND WRIGHT, M. ERIK. The Use of Content Analysis in Rorschach Interpretation: I. Differential Characteristics of Male Homosexuals. *Rorschach Res. Exch.*, 1945, 9, 169-177.
39. ELIZUR, ABRAHAM. Content analysis of the Rorschach with regard to anxiety and hostility. *Rorschach Res. Exch.*, 1949, 13, 247-284.
40. ENGLE, T. L. The use of the Harrower-Erickson Multiple Choice (Rorschach) Test in differentiating between well-adjusted and maladjusted high school pupils. *J. Educ. Psychol.*, 1946, 37, 550-556.
41. EYSENCK, H. J. A comparative study of four screening tests for neurotics. *Psychol. Bull.*, 1945, 42, 659-662.
42. EYSENCK, H. J. Screening-out the neurotic. *Lancet*, 1947, 252, 530-531.
43. FORD, M. *The application of the Rorschach Test to young children*. Univ. of Minn. Inst. Child Welf. Monog., 1946, No. 23.
44. FRANK, L. K. *Projective Methods*. Springfield, Ill.: Charles C Thomas, 1948.
45. GARFIELD, S. L. The Rorschach test in clinical diagnosis. *J. Clin. Psych.*, 1947, 3, 375-381.
46. GIBBY, ROBERT GWYN. The influence of varied experimental sets upon certain Rorschach variables. I. Stability of the intellectual variables. Microfilm of complete manuscript, Univ. Microfilms, Ann Arbor, Mich. Publ. No. 1512.
47. GOLDFARB, WILLIAM. The animal symbol in the Rorschach test and an animal association test. *Rorschach Res. Exch.*, 1945, 9, 8-22.
48. GOLDFARB, WILLIAM. Rorschach test differences between family-reared, institution-reared, and schizophrenic children. *Amer. J. Orthopsychiat.*, 1949, 19, 624-633.
49. GOLDSTEIN, KURT, AND ROTHMANN, EVA. Physiognomic phenomena in Rorschach responses. *Rorschach Res. Exch.*, 1945, 9, 1-7.
50. GOODENOUGH, FLORENCE L. The appraisal of child personality. *Psychol. Rev.*, 1949, 56, 123-131.



51. HALLOWELL, A. I. Acculturation processes and personality changes as indicated by the Rorschach technique. *Rorschach Res. Exch.*, 1942, 6, 42-50.
52. HALLOWELL, A. Popular reponses and cultural differences: An analysis based on frequencies in a group of American Indian subjects. *Rorschach Res. Exch.*, 1945, 9, 153-168.
53. HALLOWELL, A. I. The Rorschach technique in the study of personality and culture. *Amer. Anthropol.*, 1945, 47, 195-210.
54. HALPERN, F. Rorschach interpretation of the personality structure of schizophrenics who benefit from insulin therapy. *Psychiat. Quart.*, 1940, 14, 826-833.
55. HAMLIN, R. M., ALBEE, G. W., AND LELAND, E. M. Objective Rorschach "signs" for groups of normal, maladjusted and neuropsychiatric subjects. *J. Consult. Psychol.*, 1950, 14, 276-282.
56. HARROWER-ERICKSON, M. R. Large scale investigation with the Rorschach method. *J. Consult. Psychol.*, 1943, 7, 120-126.
57. HARROWER-ERICKSON, M. R., AND STEINER, M. E. *Large scale Rorschach techniques; a manual for the Group Rorschach and Multiple Choice Test*. Springfield, Ill.: Charles C Thomas, 1945.
58. HARROWER, MOLLIE R., AND STEINER, MATILDA E. *Psychodiagnostic Inkblots*. New York: Grune & Stratton, Inc., 1945.
59. HERTZ, MARGUERITE R. Rorschach norms for an adolescent age group. *Child Developm.*, 1935, 6, 69-76.
60. HERTZ, MARGUERITE R. Rorschach: twenty years after. *Psychol. Bull.*, 1942, 39, 529-572.
61. HERTZ, MARGUERITE R. Personality patterns in adolescence as portrayed by the Rorschach ink-blot method: I. The movement factors. *J. Gen. Psychol.*, 1942, 27, 119-188.
62. HERTZ, MARGUERITE R. Personality patterns in adolescence as portrayed by the Rorschach ink-blot method: III. The "Erlebnistypus" (a normative study). *J. Gen. Psychol.*, 1943, 28, 225-276.
63. HERTZ, MARGUERITE R. Personality patterns in adolescence as portrayed by the Rorschach method: IV. The "Erlebnistypus" (a typological study). *J. Gen. Psychol.*, 1943, 29, 3-45.
64. HERTZ, MARGUERITE R. Book Review: *Large scale Rorschach techniques*, by Harrower-Erickson, M. R., and Steiner, M. E. *Rorschach Res. Exch.*, 1945, 9, 46-53.
65. HERTZ, MARGUERITE R. *Frequency tables to be used in scoring responses to the Rorschach ink-blot test*. Dept. of Psychology, Western Reserve Univ., Cleveland, O., 1946. 3rd ed.
66. HERTZ, MARGUERITE R. *The Rorschach Psychogram*. Dept. of Psychology, Western Reserve Univ., Cleveland O., 1946.
67. HERTZ, MARGUERITE R. Suicidal configurations in Rorschach records. *Rorschach Research Exch. and J. of Projective Techniques*, 1948, 12, 3-58.
68. HERTZ, MARGUERITE R. Further study of "suicidal" configurations in Rorschach records. *Rorschach Res. Exch.*, 1949, 13, 44-73.

69. HERTZ, MARGUERITE R. The first international Rorschach conference. *J. Projective Techniques*, 1950, 14, 39-51.
70. HERTZ, MARGUERITE R. The Rorschach Method: science or mystery. *J. Consult. Psychol.*, 1943, 7, 67-79.
71. HERTZ, MARGUERITE R., AND BAKER, E. Personality patterns in adolescence as portrayed by the Rorschach ink-blot method: II. The color factors. *J. Gen. Psychol.*, 1943, 28, 3-61.
72. HERTZ, MARGUERITE R., AND EBERT, ELIZABETH H. The mental procedure of 6 and 8 year old children as revealed by the Rorschach ink-blot method. *Rorschach Res. Exch.*, 1944, 8, 10-30.
73. HERTZ, MARGUERITE R., ELLIS, ALBERT, AND SYMONDS, PERCIVAL M. Rorschach methods and other projective techniques. *Rev. Educ. Res.*, 1947, 17, 78-100.
74. HERTZMANN, M., AND MARGULIES, H. Developmental changes as reflected in Rorschach test responses. *J. Gen. Psychol.*, 1943, 62, 189-215.
75. HERTZMAN, MAX, AND PEARCE, JANE. The personal meaning of the human figure in the Rorschach *Psychiatry*, 1947, 10, 413-422.
76. HSÜ, E. H. The Rorschach responses and factor analysis. *J. Gen. Psychol.*, 1947, 37, 129-138.
77. HUGHES, ROBERT M. Rorschach signs for the diagnosis of organic pathology. *Rorschach Res. Exch.*, 1948, 12, 165-167.
78. HUTT, MAX L., AND SHOR, J. Rationale for routine Rorschach "Testing-the-Limits." *Rorschach Res. Exch.*, 1946, 10, 70-76.
79. JENSEN, M. B., AND ROTTER, J. B. The validity of the multiple choice Rorschach test in officer candidate selection. *Psychol. Bull.*, 1945, 42, 182-185.
80. JOEL, WALTHER. The interpersonal equation in projective methods. *Rorschach Res. Exch.*, 1949, 13, 479-482.
81. JOLLES, ISAAC. A study of mental deficiency by the Rorschach technique. *Amer. J. Ment. Def.*, 1947, 52, 37-42.
82. JOLLES, ISAAC. The diagnostic implications of Rorschach's Test in case studies of mental defectives. *Genet. Psychol. Monogr.*, 1947, 36, 89-197.
83. KABACK, G. R. *Vocational personalities: an application of the Rorschach group method*. Teach. Coll. Contr. Educ., 1946, no. 924.
84. KAY, L. W., AND VORHAUS, P. G. Rorschach reactions in early childhood. Part II. Intellectual aspects of personality development. *Rorschach Res. Exch.*, 1943, 7, 71-77.
85. KELLMAN, SAMUEL. A proposed revision of the multiple-choice Rorschach: theoretical and methodical problems. *Rorschach Res. Exch.*, 1949, 13, 244.
86. KIMBLE, G. A. Social influence on Rorschach records. *J. Abnorm. Soc. Psychol.*, 1945, 40, 89-93.
87. KELLEY, D. M., MARGULIES, H., AND BARRERA, S. E. The stability of the Rorschach method as demonstrated in electric convulsive therapy cases. *Rorschach Res. Exch.*, 1940, 5, 35-43.
88. KLOFFER, BRUNO, AND DAVIDSON, HELEN H. *The Rorschach Technique*. 1946 Supplement, New York: World Book Co.



89. KLOPPER, B., AND KELLEY, D. M. *The Rorschach Technique*. Yonkers-on-Hudson: World Book Company, 1942.
90. KRUGMAN, J. I. A clinical validation of the Rorschach with problem children. *Rorschach Res. Exch.*, 1942, 6, 61-70.
91. KRUGMAN, M. Psychosomatic study of fifty stuttering children. Round table. IV. Rorschach study. *Amer. J. Orthopsychiat.*, 1946, 16, 127-133.
92. KRUGMAN, M. *The third mental measurements yearbook*, 1949, pp. 132-133. Ed: Buros, O. K. New Brunswick, N.J.: Rutgers Univ. Press.
93. KURTZ, ALBERT K. A research test of the Rorschach test. *Personnel Psychol.*, 1948, 1, 41-51.
94. LANE, BARBARA M. A validation test of the Rorschach movement interpretation. *Amer. J. Orthopsychiat.*, 1948, 18, 292-296.
95. LAWSHE, C. H., JR., AND FORSTER, MAX H. Studies in projective techniques: I. The reliability of a multiple choice group Rorschach test. *J. Appl. Psychol.*, 1947, 31, 199-211.
96. LAZARUS, RICHARD S. The influence of color on the protocol of the Rorschach Test. *J. Abnorm. Soc. Psychol.*, 1949, 44, 506-515.
97. LEVINE, K. N., GRASSI, J. R., AND GERSON, M. J. Hypnotically induced mood changes in the verbal and graphic Rorschach: a case study. Part II: The response records. *Rorschach Res. Exch.*, 1944, 8, 104-124.
98. LINDNER, ROBERT M. Analysis of the Rorschach test by content. *J. Clin. Psychopath.*, 1947, 8, 707-719.
99. LUCHINS, ABRAHAM S. Situational and attitudinal influences on Rorschach responses. *Amer. J. Psychiat.*, 1947, 103, 780-784.
100. MACFARLANE, J. W. Problems of validation inherent in projective methods. *Amer. J. Orthopsychiat.*, 1942, 12, 405-411.
101. MALAMUD, RACHEL F., AND MALAMUD, DANIEL. The validity of the Amplified Multiple Choice Rorschach as a screening device. *J. Consult. Psychol.*, 1945, 9, 224.
102. MALAMUD, R. F., AND MALAMUD, D. I. The multiple choice Rorschach: a critical examination of its scoring system. *J. Psychol.*, 1946, 21, 237-242.
103. MARGULIES, H. *Rorschach responses of successful and unsuccessful students*. New York: Arch. Psychol., 1942, No. 271, p. 61.
104. MCCANDLESS, BOYD R. The Rorschach as a predictor of academic success. *J. Appl. Psychol.*, 1949, 33, 43-50.
105. MCCLELLAND, D. C., AND ATKINSON, J. W. The relation of the intensity of a need to the amount of perceptual distortion. *J. Psychol.*, 1947, 25, 205-222.
106. MCFATE, MARGUERITE Q., AND ORR, FRANCES G. Through adolescence with the Rorschach. *Rorschach Res. Exch.*, 1949, 13, 302-319.
107. MELTZER, H. Personality differences between stuttering and non-stuttering children as indicated by the Rorschach test. *J. Psychol.*, 1944, 17, 39-59.
108. MIALE, F. R., AND HARROWER-ERICKSON, M. R. Personality structure in the psychoneuroses. *Rorschach Res. Exch.*, 1940, 4, 71-74.
109. MILNER, B., AND MOREAULT, L. Etude du Test. Rorschach en relation au réflexe psychogalvanique. *Bull. Canad. Psychol. Assn.*, 1945, 5, 80 (abstract).

110. MILTON, E. OHMER, JR. The influence on varied experimental sets upon certain Rorschach variables: II. Stability of the human movement variable. Microfilm Abstr., 1950, 10(1), 127-128. Ph.D. thesis, 1950, U. Michigan.
111. MONTALTO, F. D. An application of the Group Rorschach technique to the problem of achievement in college. *J. Clin. Psychol.*, 1946, 2, 254-260.
112. MORRIS, WOODROW W. Prognostic possibilities of the Rorschach method in metrazol therapy. *Amer. J. Psychiat.*, 1943, 100, 222-230.
113. MORRIS, WOODROW WILBERT. The prediction of personality attributes by means of the Rorschach method. Microfilm Abstr., 1950, 9(3), 176-177. Ph.D. thesis, 1949, U. Michigan.
114. MUENCH, GEORGE A. An evaluation of non-directive psychotherapy. *Appl. Psychol. Monogr.*, 1947, No. 13, 168 pp.
115. MUNROE, RUTH L. Prediction of the adjustment and academic performance of college students by a modification of the Rorschach method. *Appl. Psychol. Monographs*, 1945, No. 7.
116. MUNROE, RUTH L. The use of projective methods in group testing. *J. Consult. Psychol.*, 1948, 12, 8-15.
117. PATTERSON, M., AND MAGAW, D. C. An investigation of the validity of the Rorschach technique as applied to mentally defective problem children. *Proc. Amer. Ass. Ment. Def.*, 1938, 43, 179-185.
118. PIOTROWSKI, Z. The Rorschach ink-blot method in organic disturbances of the central nervous system. *J. Nerv. Ment. Dis.*, 1937, 86, 525-537.
119. PIOTROWSKI, Z. A simple experimental device for the prediction of outcome of insulin treatment in schizophrenia. *Psychiat. Quart.*, 1940, 14, 267-273.
120. PIOTROWSKI, Z. A. The Rorschach method as a prognostic aid in the insulin shock treatment of schizophrenics. *Psychiat. Quart.*, 1941, 15, 807-822.
121. PIOTROWSKI, Z. A. The modifiability of personality as revealed by the Rorschach method: methodological considerations. *Rorschach Res. Exch.*, 1942, 6, 160-167.
122. PIOTROWSKI, Z. A. Tentative Rorschach formulae for educational and vocational guidance in adolescence. *Rorschach Res. Exch.*, 1943, 7, 16-27.
123. PIOTROWSKI, Z. A. Use of the Rorschach in vocational selection. *J. Consult Psychol.*, 1943, 7, 97-102.
124. PIOTROWSKI, Z. A. Experimental psychological diagnosis of mild forms of schizophrenia. *Rorschach Res. Exch.*, 1945, 9, 189-200.
125. PIOTROWSKI, Z. A., CANDEE, B. BALINSKY, B., HOLTZBERG, S., VON ARNOLD, B. Rorschach signs in the selection of outstanding young male mechanical workers. *J. Psychol.*, 1944, 18, 131-150.
126. POSTMAN, L., BRUNER, J. S., AND MCGINNIES, E. Personal values as selective factors in perception. *J. Abnorm. Soc. Psychol.*, 1948, 43, 142-154.
127. RABIN, A. I. Statistical problems involved in Rorschach patterning. *J. Clin. Psychol.*, 1950, 7, 19-21.
128. RABIN, A. I., AND BECK, S. J. Genetic aspects of some Rorschach factors. Annual meeting of Amer. Orthopsychiat. Assn., 1949.



129. RAPAPORT, D., GILL, M., AND SCHAFER, R. *Diagnostic psychological testing; the theory, statistical evaluation and diagnostic application of a battery of tests.* Vol. II. Chicago: Year Book Publishers, 1946.
130. REITZELL, JEANNE MANNHEIM. A comparative study of hysterics, homosexuals and alcoholics using content analysis of Rorschach responses. *Rorschach Res. Exch.*, 1949, 13, 127-141.
131. ROBB, R. W., KOVITZ, B., AND RAPAPORT, D. Histamine in the treatment of psychosis; a psychiatric and objective psychological study. *Amer. J. Psychiat.*, 1940, 97, 601-610.
132. ROCKWELL, F. V., WELCH, L., KUBIS, J., AND FISICHELLI, V. Changes in palmar skin resistance during the Rorschach test. I. Color shock and psychoneurotic reactions. *Monthly Rev. Psychiat. Neurol.*, 1947, 113, 129-152.
133. ROCKWELL, FRED V., WELCH, LIVINGSTON; KUBIS, JOSEPH, AND FISICHELLI, VINCENT. Changes in palmar skin resistance during the Rorschach test. II. The effect of repetition with color removed. *Mschr. Psychiat. Neurol.*, 1948, 116, 321-345.
134. ROE, ANNE, A Rorschach Study of a group of scientists and technicians. *J. Consult. Psychol.*, 1945, 14, 317-327.
135. ROE, ANNE. Artists and their work. *J. Personality*, 1946, 15, 1-40.
136. ROE, ANNE. Personality and vocation. *Trans. N.Y. Acad. Sci.*, 1947, 9, 257-267.
137. ROE, ANNE. Psychological examinations of eminent biologists. *J. Consult. Psychol.*, 1949, 13, 225-246.
138. ROSS, W. DONALD. Relation between Rorschach interpretations and clinical diagnosis. *J. Proj. Techniques*, 1950, 14, 5-14.
139. ROSS, W. D., FERGUSON, G. A., AND CHALKE, F. C. R. The group Rorschach in officer selection. *Bull. Canad. Psychol. Ass.*, 1945, 5, 84-86.
140. ROSS, W. D., AND ROSS, S. Some Rorschach ratings of clinical value. *Rorschach Res. Exch.*, 1944, 8, 1-9.
141. ROTTER, JULIAN B. The present status of the Rorschach in clinical and experimental procedures. *J. Personality*, 1948, 16, 304-311.
142. RUESCH, J., AND FINESINGER, J. E. The relation of the Rorschach color response to the use of color in drawings. *Psychosom. Med.*, 1941, 3, 370-388.
143. RUST, RALPH MASON. Some correlates of the movement response. *J. Personality*, 1948, 16, 369-401.
144. SAPPENFELD, BART R., AND BUKER, SAMUEL L. Validity of the Rorschach 8-9-10 per cent as an indicator of responsiveness to color. *J. Consult. Psychol.*, 1949, 13, 268-271.
145. SARASON, SEYMOUR B., AND SARASON, ESTHER KROOP. The discriminatory value of a test pattern in the high grade familial defective. *J. Clin. Psychol.*, 1946, 2, 38-49.
146. SARASON, SEYMOUR B., AND SARASON, ESTHER KROOP. The discriminatory value of a test pattern with cerebral palsied defective children. *J. Clin. Psychol.*, 1947, 3, 141-146.
147. SCHACHTEL, ERNEST G. The dynamic perception and the symbolism of form. *Psychiatry*, 1941, 4, 79-96.

148. SCHACHTEL, ERNEST G. On color and affect. *Psychiatry*, 1943, 6, 393-409.
149. SCHACHTEL, E. G. Subjective definitions of the Rorschach test situation and their effect on test performance. Contributions to an understanding of Rorschach's test, III. *Psychiatry*, 1945, 8, 419-448.
150. SCHACHTEL, ERNEST G. Projection and its relation to character attitudes and creativity in the kinesthetic responses. *Psychiatry*, 1950, 13, 69-100.
151. SHAW, BARRIE. "Sex populars" in the Rorschach test. *J. Abnorm. Soc. Psychol.*, 1948, 43, 466-470.
152. SHOEMAKER, H. A., AND ROHRER, J. H. Relationship between success in the study of medicine and certain psychological and personal data. *J. Ass. Amer. Med. Coll.*, 1948, 23, 190-201.
153. SIEGEL, MIRIAM G. The diagnostic and prognostic validity of the Rorschach test in a child guidance clinic. *Amer. J. Orthopsychiat.*, 1948, 18, 119-133.
154. SIIPOLA, ELSA M. The influence of color on reaction to ink-blot. *J. Personality*, 1950, 18, 358-383.
155. SLOAN, WILLIAM. Mental deficiency as a symptom of personality disturbance. *Amer. J. Mental Def.*, 1947, 52, 31.
156. SPRINGER, N. NORTON. The validity of the multiple choice group Rorschach test in the screening of naval personnel. *J. Gen. Psychol.*, 1946, 35, 27-32.
157. STAINBROOK, E. The Rorschach description of immediate post-convulsive mental function. *Character and Pers.*, 1944, 12, 302-322.
158. STAINBROOK, E., AND SIEGEL, P. S. A comparative group Rorschach study of Southern Negro and white high school and college students. *J. Psychol.*, 1944, 17, 107-115.
159. STAVRIANOS, B. An investigation of sex differences in children as revealed by the Rorschach method. *Rorschach Res. Exch.*, 1942, 6, 168-175.
160. STEIN, MORRIS I. Personality factors of Rorschach responses. *Rorschach Res. Exch.*, 1949, 13, 355-413.
161. STEINER, MATILDA E. The use of the Rorschach method in industry. *Rorschach Res. Exch.*, 1947, 11, 46-52.
162. STEINER, MATILDA E. *The psychologist in industry*. Springfield, Ill.: Charles C Thomas, 1949.
163. STEPHENSEN, WILLIAM. A statistical approach to typology. *J. Clin. Psychol.* (Monograph Supplement) No. 7, 1950, 26-38.
164. SWIFT, J. W. Matchings of teachers' descriptions and Rorschach analyses of preschool children. *Child Developm.*, 1944, 15, 217-224.
165. SWIFT, J. W. Reliability of Rorschach scoring categories with preschool children. *Child Developm.*, 1944, 15, 207-216.
166. SWIFT, J. W. Relation of behavioral and Rorschach measures of insecurity in preschool children. *J. Clin. Psychol.*, 1945, 1, 196-205.
167. SWIFT, J. W. Rorschach responses of eighty-two preschool children. *Rorschach Res. Exch.*, 1945, 9, 74-84.
168. THOMPSON, GRACE M. College grades and the group Rorschach. *J. Appl. Psychol.*, 1948, 32, 398-407.



169. THOMPSON, GRACE M. MMPI correlates of certain movement responses in the group Rorschach of two college samples. *J. Consult. Psychol.*, 1948, 12, 379-383.
170. THURSTONE, L. L. The Rorschach in psychological science. *J. Abnorm. Soc. Psychol.*, 1948, 43, 471-475.
171. VORHAUS, P. G. Rorschach reactions in early childhood. Part III. Content and details in pre-school records. *Rorschach Res. Exch.*, 1944, 8, 71-91.
172. VORHAUS, P. G. Non-reading as an expression of resistance. *Rorschach Res. Exch.*, 1946, 10, 60-69.
173. WALLEN, RICHARD. The nature of color shock. *J. Abnorm. Soc. Psychol.*, 1948, 43, 346-356.
174. WEISSKOPF, E. A. The influence of the time factor on Rorschach Performance. *Rorschach Res. Exch.*, 1942, 6, 128-136.
175. WERNER, H. Perceptual behavior of brain injured, mentally defective children: an experimental study by means of the Rorschach technique. *Genet. Psychol. Monogr.*, 1945, 31, 51-110.
176. WHEELER, WILLIAM MARSHALL. An analysis of Rorschach indices of male homosexuality. *Rorschach Res. Exch.*, 1949, 13, 97-126.
177. WHEELER, W., BUHLER, C., GRAYSON, H., MEYER, M., WESLEY, S., AND LEFEVER, D. Symposium on a Basic Rorschach Score. *Rorschach Res. Exch.*, 1949, 13, 6-24.
178. WILKINS, WALTER L., AND ADAMS, AUSTIN J. The use of the Rorschach test under hypnosis and under sodium amytol in military psychiatry. *J. Gen. Psychol.*, 1947, 36, 131-138.
179. WILLIAMS, MEYER. An experimental study of intellectual control under stress and associated Rorschach factors. *J. Consult. Psychol.*, 1947, 11, 21-29.
180. WINFIELD, M. C. The use of the Harrower-Erickson Multiple Choice Rorschach Test with a selected group of women in military service. *J. Appl. Psychol.*, 1946, 30, 481-487.
181. WISHNER, JULIUS. Rorschach intellectual indicators in neurotics. *Amer. J. Orthopsychiat.*, 1948, 18, 265-279.
182. WITTENBORN, J. R. Certain Rorschach response categories and mental abilities. *J. Appl. Psychol.*, 1949, 33, 330-338.
183. WITTENBORN, J. R. A factor analysis of discrete responses to the Rorschach ink blots. *J. Consult. Psychol.*, 1949, 13, 335-340.
184. WITTENBORN, J. R. Statistical tests of certain Rorschach assumptions: analysis of discrete responses. *J. Consult. Psychol.*, 1949, 13, 257-267.
185. WITTENBORN, J. R. Statistical tests of certain Rorschach assumptions: The internal consistency of scoring categories. *J. Consult. Psychol.*, 1950, 14, 1-19.
186. WITTENBORN, J. R., AND SARASON, SEYMOUR B. Exceptions to certain Rorschach criteria of pathology. *J. Consult. Psychol.*, 1949, 13, 21-27.
187. WITTON, C. L., HUNT, W. A., AND OLDER, H. J. The use of the Multiple Choice Group Rorschach Test in military screening. *J. Psychol.*, 1944, 17, 91-94.

188. ZUBIN, J. Introduction. The problems of quantification and objectification in personality: A symposium. *J. Personality*, 1948, 17, 141-145.
189. ZUBIN, J. Personality research and psychopathology as related to clinical practice. Clinical practice and personality theory: A symposium. *J. Abnorm. and Social Psychology*, 1949, 44, 12-21.
190. ZUBIN, J. Introduction, Symposium on statistics for the clinician. *J. Clin. Psychol.*, 1950, 6, 1-6.
191. ZUBIN, J., CHUTE, E., AND VENIAR, S. Psychometric scales for scoring Rorschach test responses. *Character and Pers.*, 1943, 11, 277-301.
192. ZUBIN, J., AND YOUNG, K. M. *Manual of Projective and Cognate Techniques*, Madison, Wis., College Typing Co., 1948 (mimeographed).
193. ZULLIGER, HANS. Personality dynamics as revealed in the Rorschach of a 15 year old girl. *J. Projective Techniques*, 1950, 14, 52-60.



# NAME INDEX

- Abel, T. M., 48, 88, 361,  
 383, 406, 420  
 Abt, L. E., 413, 420  
 Ackerman, N. W., 48, 70,  
 73, 90  
 Adams, A. J., 103, 143, 231,  
 405, 429  
 Ahman, J. S., 16  
 Ainsworth, M., 313  
 Aita, J. A., 406, 420  
 Albee, G. W., 211, 338, 345,  
 410, 421  
 Albrecht, R., 338, 344  
 Allen, F. H., 90, 91  
 Allport, F. H., 85, 97  
 Allport, G. W., 31, 48, 56,  
 65, 79, 80, 97, 419, 420  
 Altus, W. D., 407, 420  
 Amen, E. W., 67, 94  
 Ames, L. B., 313, 344  
 Anastasi, Anne, 16, 88, 270,  
 289  
 Anderson, H. H., 48  
 Anderson, I., 413, 420  
 Anderson, J. E., 82  
 Anderson, R. G., 413, 420  
 Armitage, S. G., 406, 420  
 Atkinson, J. W., 393, 425  
  
 Baker, L. M., 397, 403, 420,  
 424  
 Baldwin, A. L., 66, 98  
 Balinsky, B., 162, 406, 426  
 Balken, E. R., 51, 55, 65, 81,  
 82, 87, 94, 98  
 Barker, R. G., 49  
 Barnhart, E. N., 88  
 Barrere, S., 379, 385, 402,  
 424  
 Barron, F. X., 335, 344  
 Barry, J. R., 338, 344  
 Bartlett, F. C., 61, 98  
 Baruch, D. W., 73, 90  
  
 Bateson, G., 33, 41  
 Baughman, E. E., 211, 333,  
 339, 344  
 Bechtoldt, H. P., 266, 289  
 Beck, E., 89  
 Beck, L. W., 276, 289  
 Beck, S. J., 49, 62, 84, 92,  
 142, 162, 200, 207, 211,  
 217, 229, 295, 308, 365,  
 383, 394, 396, 420  
 Beers, C., 66, 98  
 Behn-Eschenburg, H., 295,  
 297, 308  
 Bell, J. E., 394, 405, 421  
 Bellak, L., 73, 94, 102, 142  
 Bender, C., 49  
 Bender, Lauretta, 59, 68,  
 69, 70, 88, 89, 90, 95  
 Benedict, R., 33  
 Benjamin, A. C., 98  
 Benjamin, J. D., 92, 223,  
 229, 334, 344  
 Bennet, Georgia, 94  
 Berger, S., 314  
 Bergman, M. S., 103, 142,  
 220, 229, 396, 401, 421  
 Berkowitz, M., 211  
 Beyer, E., 87  
 Binet, A., 61, 73, 98  
 Blair, W. R. N., 289  
 Blake, R. R., 225, 229  
 Blanchard, P., 91  
 Bleuler, E., 63, 98, 231, 297,  
 309  
 Blumenthal, N. T., 43  
 Blyth, D. D., 338, 344  
 Bochner, R., 62, 92, 116,  
 142, 394, 398, 421  
 Booth, G. C., 49, 87  
 Bordin, R., 89  
 Boynton, P. L., 98  
 Bradway, K. P., 409, 421  
 Braithwaite, R. B., 276, 289  
  
 Brenman, M., 85, 92  
 Bridgman, P. W., 60, 98  
 Brittain, H. L., 98  
 Brosin, H. W., 216, 229  
 Brower, D., 16, 403, 421  
 Brown, E., 87  
 Brown, J. F., 87, 98  
 Brown, R. R., 353, 384  
 Brozek, J., 413, 421  
 Bruner, J. S., 393, 421  
 Brussel, J., 115, 142  
 Bucher, J., 97  
 Buhler, C., 73, 96, 116, 142,  
 356, 365, 384, 395, 421  
 Buhler, K., 356, 384, 421  
 Buker, S. L., 404, 427  
 Burke, C. J., 211  
 Burks, B. S., 31  
 Burnham, C. A., 407, 421  
 Burr, H. S., 37  
  
 Cameron, N., 49, 50  
 Cameron, W. M., 71, 90  
 Candee, B., 406, 426  
 Carnap, R., 276, 289  
 Carr, A. C., 338, 344, 410,  
 421  
 Cartwright, D., 80, 98  
 Cattell, R. B., 63, 421  
 Chalke, F. L. R., 386  
 Challman, R. C., 162, 264,  
 412, 421  
 Chapman, D. W., 234, 263  
 Cheshire, L., 162  
 Chidester, L., 88  
 Child, I. L., 289  
 Christenson, J. A., 94  
 Christiansen, C., 356, 385  
 Chute, E., 94, 396, 430  
 Chyatte, C., 280, 289  
 Clapp, H., 93  
 Clark, J. H., 407, 421  
 Clark, L. P., 61, 98

- Cochran, W. G., 352, 384  
 Cofer, C. N., 103, 142, 401, 421  
 Coghill, G. E., 43  
 Cohen, B. D., 219, 230  
 Conklin, E. S., 61, 98  
 Conn, J. H., 50, 70, 90  
 Cottrell, L. S., 16  
 Cowell, B., 89  
 Cox, K. J., 335, 344, 413, 422  
 Cronbach, L. J., 15, 18, 25, 26, 211, 235, 263, 264, 289, 334, 345, 347, 384, 401, 422, 410  
 Crossland, H. R., 64, 98  
 Cureton, E. E., 290  
 Curran, J., 50, 89  
  
 Damin, D. E., 290  
 Dana, R. H., 15, 310, 313  
 Davidson, H. H., 336, 345, 375, 384, 396, 405, 416, 422, 424  
 Davis, F. P., 86, 98  
 Dembo, T., 49  
 Despert, J. L., 50, 68, 70, 90, 96  
 Dorken, H. J., 335, 345  
 Dubin, S. S., 73, 96  
 Due, F. O., 396, 422  
 Duff, I. F., 98  
 Dunbar, H. F., 44  
 du Nouy, P. L., 37  
  
 Earl, C. J. C., 225, 229  
 Ebaugh, F. G., 92, 334, 344  
 Ebert, E. H., 397, 424  
 Edwards, A. L., 384  
 Eichler, R. M., 315, 322, 333, 345  
 Elizur, A., 24, 26, 337, 345, 396, 422  
 Ellis, A., 405, 424  
 Engle, T. L., 162, 414, 422  
 Enke, W., 295, 308  
 Erikson, E. H., 50, 59, 69, 70, 76, 90  
 Eysenck, H. J., 273, 290, 414, 422  
  
 Fearing, F., 313  
 Feigenbaum, D., 55, 87  
 Feigl, H., 276, 290  
 Ferguson, G. A., 386  
 Festinger, L., 362, 384  
 Finesinger, J. E., 403, 427  
 Fisher, M. S., 85, 97  
 Fisichelli, V., 222, 230, 403, 427  
 Fiske, D. W., 211, 281, 290  
 Fite, M. D., 48, 50, 73, 96  
 Fleming, J., 88  
 Foley, J. P., 88  
 Ford, M., 397, 418, 420  
 Forster, E. M., 162, 414, 425  
  
 Fosberg, I. A., 78, 92, 102, 142, 380, 384  
 Foulds, G., 96  
 Frank, L. K., 14, 31, 32, 35, 53, 54, 59, 60, 65, 66, 68, 69, 79, 80, 83, 87, 92, 216, 229, 233, 263, 417, 422  
 Franz, J. G., 89  
 Freeman, H., 366, 384  
 Freidman, H., 203, 211  
 French, J. R., 80, 98  
 Freud, A., 70, 91, 98  
 Freud, S., 54, 55, 56, 74, 98  
 Fries, M. E., 91  
 Fromm, E. O., 216, 229  
  
 Galton, F., 61, 63, 98  
 Gann, E., 353, 384  
 Garfield, S. L., 407, 422  
 Garrett, H. E., 98  
 Gaylord, R. H., 290  
 Gerard, M., 91  
 Gerson, M. T., 143, 220, 230, 405, 425  
 Gibby, R. G., 315, 322, 394, 422  
 Gibson, J. J., 342, 345  
 Gitelson, M., 50, 91  
 Goldensohn, L. N., 92  
 Goldfarb, W. A., 351, 353, 384, 396, 405, 422  
 Goldstein, K., 396, 424  
 Goodenough, F. L., 265, 290, 422  
 Goodman, C. C., 393, 421  
 Gordon, R. G., 297, 308  
 Gough, H. G., 280, 290  
 Graham, H., 103, 142, 221, 229  
 Graham, V. T., 223, 229, 405, 421  
 Granger, G. W., 22, 26  
 Grant, D. A., 98  
 Grant, M. Q., 335, 345  
 Grassi, J. R., 143, 220, 230, 405, 425  
 Grayson, H., 429  
 Green, G. H., 61, 98  
 Griffiths, R. A., 50, 63, 99  
 Grings, W. W. X., 65, 68, 95  
 Guetzkow, H., 413, 421  
 Guilford, J. P., 200, 265, 272, 290, 334, 345, 357, 367, 384  
 Guirdham, A., 229  
 Gulliksen, H., 9, 16, 265, 274, 290  
 Gustav, A., 200, 357, 385  
 Guttman, E., 221, 229  
  
 Haggard, E. A., 96  
 Hall, C. S., 142  
 Hall, G. S., 61, 99  
 Hallowell, A. I., 396, 418, 423  
 Halpern, F., 62, 92, 116, 142, 223, 229, 394, 421, 423  
  
 Hamlin, R. M., 211, 338, 345, 410, 421, 423  
 Hammond, K. R., 23, 26, 335, 345  
 Hanfman, E., 50  
 Harms, E., 88  
 Harris, J. S., 403, 420  
 Harris, R. E., 356, 385  
 Harris, T. M., 385  
 Harrison, R., 58, 62, 70, 75, 79, 81, 83, 84, 94, 216, 229, 263  
 Harrower-Erickson, M. R., 92, 150, 162, 177, 200, 211, 402, 412, 423  
 Harvey, N. A., 99  
 Hathaway, S. R., 291  
 Heider, F., 142  
 Hempel, C. G., 276, 280, 290  
 Hertz, M. R., 15, 50, 76, 78, 80, 84, 92, 93, 205, 207, 208, 211, 217, 230, 232, 238, 239, 263, 295, 309, 310, 313, 345, 352, 353, 385, 391, 397, 409, 411, 423, 424  
 Hertzman, M., 93, 142, 353, 356, 365, 373, 385, 396, 407, 424  
 Hibbard, A., 55  
 Higginbotham, S. A., 94  
 Hirt, M., 2  
 Hitch, K. S., 115, 142  
 Holly, S. B., 89  
 Holmer, P., 50, 91  
 Holt, R., 23, 25, 26, 313  
 Holtzberg, S., 335, 346, 406, 426  
 Holzberg, J. D., 203, 212  
 Homburger, E., 70, 91  
 Horowicz, E. H., 67, 72  
 Horowitz, R. E., 44, 50, 87, 96  
 Horst, P., 290  
 Hovey, H. B., 280, 290  
 Hsu, E. H., 142, 409, 424  
 Hughes, R. M., 406, 408, 424  
 Hull, C. L., 99  
 Hunt, W. A., 102, 142, 414, 429  
 Hunter, M., 50, 234, 263  
 Hutt, M. L., 315, 322, 394, 424  
  
 Ives, V., 335, 345  
  
 Jacob, Z., 142  
 Jaensch, W., 46, 50  
 Jastrow, J., 63  
 Jenkins, R. L., 89, 265, 290  
 Jennings, H., 51  
 Jensen, M. B., 162, 414, 424  
 Jersild, A., 48  
 Joel, W., 401, 424  
 Johnson, W., 65  
 Jolles, I., 406, 424



- Jones, M. C., 31  
 Jung, C. G., 63  
  
 Kaback, G. R., 351, 359, 378, 385, 413, 424  
 Kadinsky, D., 142  
 Kamman, G. R., 142  
 Kanner, L., 91  
 Kantor, J. H., 60  
 Kaplan, A., 93, 281, 290  
 Kaplan, B., 314  
 Kasanin, J., 50  
 Kay, L. W., 397, 424  
 Keiser, S., 49  
 Kelley, D. M., 51, 62, 93, 115, 142, 200, 211, 230, 238, 241, 255, 263, 345, 379, 385, 394, 410, 424, 425  
 Kellman, S., 414, 424  
 Kelly, E. L., 281, 290  
 Kelly, G. A., 96  
 Kent, G. H., 35, 63, 64  
 Kerr, M., 59, 78, 96  
 Keys, A., 413, 421  
 Kimble, G. A., 315, 322, 394, 424  
 Kisker, G. W., 93  
 Klatskin, E. H., 315, 322  
 Klebanoff, S. B., 95  
 Klein, M., 70, 71, 91  
 Klopfer, B., 51, 52, 62, 93, 112, 142, 200, 211, 217, 230, 238, 255, 263, 313, 394, 425  
 Klopfer, W., 313  
 Klüver, H., 51  
 Knauss, J., 335, 345  
 Kneale, W., 276, 291  
 Knopf, I. J., 15, 202, 211  
 Kobler, F. J., 202, 211  
 Kovitz, B., 230, 405, 427  
 Kral, A., 335, 345  
 Krug, O., 52, 378  
 Krugman, J. E., 93, 233, 263, 350, 383, 402, 408, 425  
 Krugman, M., 61, 80, 93, 242, 263, 353, 385, 425  
 Kubis, J., 222, 230, 403, 427  
 Kurtz, A. K., 357, 385, 406, 413, 425  
  
 Laiselle, R. H., 25, 27  
 Lane, B. M., 405, 425  
 Lawshe, C. H., 162, 414, 425  
 Lazarus, R. S., 404, 425  
 Learned, J., 313, 344  
 Leavitt, H. C., 103, 142, 221, 229, 405, 421  
 Lefever, D., 142, 356, 365, 384, 421  
 Lehrmann, P. R., 61  
 Leland, E. M., 211, 410, 421  
 Lerner, E., 69, 72, 79, 87  
 Leverett, H. M., 385  
 Levine, J., 211  
 Levine, K. N., 103, 143, 220, 230, 405, 425  
  
 Levy, D., 51, 69, 70, 91, 386  
 Levy, J., 51, 88  
 Lewin, K., 4, 16, 35, 49, 51, 56, 66  
 Lewis, D., 211  
 Lewis, N. D. C., 89  
 Lillian, K. K., 230  
 Linder, R. M., 143, 200, 396, 425  
 Lindquist, E. F., 85, 114, 143, 291, 385  
 Lippman, H. S., 91  
 Liss, E., 51, 71, 91  
 Loevinger, J., 9, 16  
 Lonstein, M., 212  
 Loosli-Usteri, M., 295, 309  
 Lopfe, A., 295, 309  
 Lord, E. E., 9, 14, 16, 101, 315, 322, 338, 345  
 Lowell, F., 64, 100  
 Lowenfield, M., 51, 59, 91, 97  
 Lowry, L., 91  
 Lucas, C. M., 275, 291  
 Luchins, A. S., 103, 143, 394, 425  
 Lugoff, L. S., 64, 99  
 Lunberg, B. A., 16  
 Lyle, J., 89  
  
 McCall, R. J., 335, 345  
 McCandless, B. R., 269, 291, 363, 368, 385, 425  
 McClelland, D. C., 393, 425  
 McClosky, H., 290  
 MacCorquodale, K., 268, 291  
 MacFarlane, J. W., 79, 81, 85, 87, 216, 230, 233, 263, 291, 394, 425  
 McFate, M. Q., 395, 397, 425  
 McGinnies, E., 393, 425  
 McIntosh, J. R., 89  
 McKinely, J. C., 291  
 McNemar, Q., 353, 385  
 Magaw, D. C., 408, 426  
 Malamud, D., 414, 425  
 Malamud, R. F., 414, 425  
 Margulies, H., 353, 365, 373, 385, 397, 425  
 Marquidt, S., 73, 97  
 Maslow, A. H., 56, 57, 58, 87  
 Masserman, J. H., 51, 65, 94, 95  
 Mathers, V. G., 52  
 Mayer, A. M., 71, 91  
 Mayer, E. B., 71, 91  
 Mead, M., 33  
 Meehl, P. E., 15, 16, 264, 268, 291  
 Meltzer, H., 335, 385, 405, 425  
 Metraux, R. W., 313, 344  
 Meyer, M., 429  
 Miale, F. R., 93, 211, 406, 425  
 Milner, B., 403, 425  
  
 Milton, E. O., 322, 394, 426  
 Mira, L. E., 68, 97, 297, 309  
 Mittelmann, B., 87  
 Moellenhoff, F., 87  
 Montalto, F. D., 356, 385, 409, 426  
 Mooney, B., 17  
 Moreault, L., 403, 425  
 Moreno, J. L., 51, 68, 89  
 Morgan, C. D., 51, 62, 95  
 Morris, W. W., 224, 230, 404, 409, 426  
 Morse, M. E., 36  
 Mosier, C. I., 265, 291  
 Mosse, E. P., 89  
 Muench, G. A., 211, 345, 398, 406, 426  
 Muller, M., 295, 309  
 Munroe, R. L., 76, 81, 93, 347, 349, 361, 374, 385, 386, 406, 426  
 Munz, E., 309  
 Murphy, L. B., 44, 50, 68, 72, 79, 87, 97  
 Murray, H. A., 51, 56, 57, 58, 62, 70, 71, 72, 74, 78, 80, 86, 87, 95, 237, 263  
 Murstein, B. I., 20, 27  
  
 Nadel, A. B., 211  
 Naumberg, M., 89  
 Neidt, C. O., 16  
 Newell, H. W., 91  
 Newman, S., 52  
 Norman, R. M., 297, 308  
 Northrop, F. S., 37  
  
 Oberholzer, E., 62, 93, 295, 309  
 Oeser, O. A., 295, 309  
 Older, H. J., 414, 429  
 Orlandy, J., 20, 27, 356, 385  
 Orr, F. G., 395, 425  
  
 Palmer, J. O., 15, 232  
 Pap, A., 276, 291  
 Patterson, M., 408, 426  
 Peak, H., 265, 291  
 Pearce, J., 142, 396, 407, 424  
 Pearson, K., 242, 263  
 Peatmen, J. G., 353, 385  
 Pemberton, W., 230  
 Penrose, L. S., 377, 386  
 Pfahler, G., 295, 309  
 Piaget, J., 64  
 Pickard, P. M., 313, 332, 345  
 Pickford, R. W., 68, 89, 97  
 Piotrowski, Z., 52, 93, 143, 200, 211, 223, 225, 230, 355, 386, 406, 410, 426  
 Plant, J. S., 52  
 Porter, E. L., 52  
 Porteus, S. D., 274, 291  
 Postman, L., 396, 421, 425  
 Potter, E., 50  
 Proshansky, H. M., 73, 97  
  
 Queen, S. A., 16

- Rabin, A. I., 397, 415, 426  
 Ramzy, I., 313, 332, 345  
 Ranzoni, J. A., 335, 345  
 Rapaport, D., 87, 95, 97, 162,  
 191, 200, 216, 230, 357, 368,  
 386, 394, 405, 411, 427  
 Reichard, S., 85, 92  
 Reiman, G. W., 203, 212  
 Reitan, R. M., 406, 420  
 Reitman, F., 89  
 Reitzell, J. M., 396, 427  
 Richards, S. S., 91  
 Richardson, L. H., 353, 386  
 Richenberg, W., 88  
 Rickers-Ovsiankina, M., 353,  
 368, 382  
 Robb, R. W., 230, 405, 427  
 Rockwell, F. W., 222, 230,  
 403, 427  
 Rodnick, E. H., 95  
 Roe, A., 413, 427  
 Roemer, G. A., 297, 309  
 Rogers, L. S., 335, 345  
 Rogerson, C. H., 91  
 Rohrer, J. H., 409, 428  
 Rorschach, H., 19, 27, 62, 93,  
 111, 115, 143, 148, 162, 177,  
 200, 212, 230, 295, 304, 309,  
 331, 345  
 Rosanoff, A. J., 63, 64, 84  
 Rosenzweig, S., 52, 71, 88, 91,  
 95, 230, 239, 263  
 Ross, S., 91, 356, 386  
 Ross, W. D., 356, 373, 386,  
 401, 427  
 Rothmann, E., 396, 422  
 Rotter, J. B., 75, 95, 162, 411,  
 414, 424, 427  
 Rubenstein, B. B., 50, 239,  
 263  
 Rubin, H., 212  
 Ruesch, J., 403, 427  
 Russell, E. S., 56, 280, 291  
 Rust, R. M., 404, 427  
 Ruth, J. M., 406, 420  
  
 Saffir, M., 162  
 Samuel, E. A., 88  
 Sandler, J., 337, 345  
 Sapir, E., 52  
 Sappenfeld, B. R., 404, 426  
 Sarason, E. R., 405, 427  
 Sarason, S., 73, 95, 405, 427  
 Sarbin, T. R., 16, 230  
 Sargent, H. D., 14, 53, 68, 71,  
 91, 97, 216, 230  
 Schachtel, E. G., 102, 143,  
 394, 407, 427, 428  
 Schilder, P., 49, 90  
 Schmidle-Wachner, T., 89  
 Schmidt, H. O., 365, 386  
 Schneider, L. I., 15, 215  
 Schube, K., 89  
 Schwartz, L. A., 95  
 Sears, R. R., 54, 74, 88  
 Seashore, R. H., 58, 100  
 Seitz, C. P., 356, 385  
  
 Sellars, W. S., 276, 291  
 Sen, A. A., 337, 345  
 Sender, S., 52  
 Shakow, D., 52, 59, 68, 71,  
 91, 96  
 Shaw, B., 396, 428  
 Shaw, R. F., 52, 89  
 Shoemaker, H. A., 409, 428  
 Shor, J., 394, 424  
 Sichá, K., 313  
 Sichá, M., 313  
 Siegel, M. G., 353, 386, 409,  
 413, 428  
 Siipola, E. M., 341, 345, 404,  
 428  
 Simon, T., 61, 98  
 Simpson, G., 91  
 Skinner, B. F., 68, 96  
 Sloan, W., 405, 428  
 Slutz, M., 73, 95  
 Snedecor, G. W., 85, 100  
 Solomon, J. C., 70, 71, 91  
 Soukup, F., 295, 309  
 Southard, E. E., 65, 100  
 Spiegelman, M., 313  
 Spiker, C. C., 269, 291  
 Spoerl, D. T., 89  
 Springer, N. N., 89, 414, 428  
 Stagner, R., 31  
 Stainbrook, E., 404, 413, 428  
 Stauffacher, J., 337, 346  
 Stavrianos, B., 398, 428  
 Steil, A., 202, 211  
 Stein, L. T., 52  
 Stein, M., 393, 404, 428  
 Steiner, M., 162, 401, 412,  
 423, 428  
 Steisel, L., 219, 229  
 Stephensen, W., 419, 428  
 Stern, W., 44, 56, 67, 97  
 Stone, L., 87, 97  
 Stouffer, S., 16, 100  
 Strang, R., 88  
 Struve, K., 309  
 Swift, J., 354, 379, 386, 397,  
 428  
 Swineford, F., 386  
 Symonds, P., 67, 88, 91, 95,  
 100, 233, 263  
  
 Tallman, F., 92  
 Terwillinger, C., 313  
 Thiesen, J. W., 212  
 Thompson, Grace M., 356,  
 386, 407, 420, 428, 429  
 Thornton, G., 200, 367, 386  
 Thurstone, L., 15, 69, 79,  
 100, 162, 217, 231, 266,  
 292, 325, 400, 429  
 Tomkins, S., 78, 85, 95, 232,  
 239, 263  
 Troup, E., 52, 350, 380, 383,  
 386  
 Trussell, M., 68, 96  
 Tucker, J., 336, 345  
 Tuddenheim, R., 73, 97  
  
 Tulchin, S., 386  
 Updegraff, R., 88  
  
 Vanderveer, A., 94  
 Varendonck, J., 61, 100  
 Varvel, W., 93, 217, 231  
 Vaughn, J., 52, 378, 387  
 Verniar, E., 94, 396, 430  
 Vernon, M., 61, 100  
 Vernon, P., 48, 62, 78, 94,  
 234, 245, 263, 295, 300,  
 303, 309, 350, 379, 387  
 Vigotsky, L., 36, 41, 100  
 Vorhaus, P., 397, 405, 424,  
 429  
  
 Walder, R., 92  
 Walker, H., 387  
 Walker, R., 313, 344  
 Wallen, R., 404, 429  
 Walsworth, B., 98  
 Watkins, J., 337, 346  
 Weil, H., 309  
 Weiss, F., 92  
 Weisskopf, E., 393, 429  
 Welch, L., 222, 230, 403,  
 427  
 Wells, F., 64, 100, 309  
 Werner, H., 52, 353, 366,  
 387, 405, 429  
 Wert, J., 16  
 Wertham, F., 59, 69, 97, 231,  
 297, 309  
 Wesley, S., 429  
 Wheeler, W., 396, 407, 429  
 White, R., 56, 88  
 Wilkins, W., 103, 143, 231,  
 405, 429  
 Williams, J., 89  
 Williams, M., 221, 231, 383,  
 387, 403, 429  
 Willoughby, A., 36  
 Wilson, G., 336, 346  
 Windle, C., 335, 346  
 Winfield, M., 162, 414, 429  
 Wishner, G., 407, 429  
 Wittenborn, J., 14, 147, 163,  
 175, 176, 201, 203, 212,  
 335, 346, 408, 429  
 Wittson, C., 414, 429  
 Wolff, E., 91  
 Wolff, W., 88  
 Woltmann, A., 49, 68, 70,  
 88, 89, 90  
 Woodrow, H., 64, 100  
 Woodworth, R., 64  
 Wright, M., 396, 421  
 Wyatt, F., 95  
  
 Yates, F., 349, 387  
 Young, K., 396, 430  
 Young, R., 94  
  
 Zax, M., 25, 27  
 Zubin, J., 15, 24, 27, 84, 94,  
 98, 331, 346, 396, 419, 430  
 Zulliger, H., 402, 430



# SUBJECT INDEX

- Academic failure and success, 406-409
- Achievement of college women, 409
- Achieving students, 413
- Actuarial approach, 3
- Adolescent records, 397
- Age level regression, 405
- Age levels, Rorschach records for, 397
- Alcoholics, responses of, 396
- Anatomy responses, 210
- Animal symbol, 396
- Approach, totalistic, 83, 123
- Armed services, Rorschach in, 62
- Assessment technique, 18
- Associative blocking, 404
- Attitude, influence of, on Rorschach, 394
- Autokinetic phenomena, 217
  
- Basic Rorschach score, 380, 395, 406
- Behavior
  - disorganization, 404
  - individual, 4
  - prediction of, 3, 25
- Blind analysis, 334
  - See also* diagnosis
- Brain damage, 203, 405, 409
- Brush developmental studies, 395
  
- C' determinants, 119
- Cardiovascular activity, 403
- Cathartic responses, 43, 68
- Cerebral palsied defectives, 405
- Check list
  - approaches, 248
  - Harrower-Erickson, 150, 177, 408
  - multiple, 179
  - scores, 374
  - validation, 258
- Chi-square, use of, 361
  
- Clay modeling, 67
- Clinical
  - diagnosis, 62
  - judgment, 401
  - method, 333
  - treatment of data, 348
- Cloud pictures, 67
- Collateral experimental approach, 216
- College students, adjustment of, 409
- Color, 12, 163, 200, 403, 407-408
  - affective value of, 403
  - associative responsiveness of, 403
  - autonomic responsiveness, 404
  - classes of, 173
  - combining responses of, 199
  - deterioration, 415
  - differentiated from movement, 163, 199
  - emotional instability, 21
  - emotional reactions, 404
  - human movement, 173, 184
  - interrelations among responses, 180
  - related to perceptual control, 197
  - similarities among, 173
  - size of area, 197
  - total score, 183, 187
- Color shock, 12, 341, 403
  - affective responsiveness, 404
  - contour of blot area, 341
- Composite score, 379
  - discrimination by, 374
- Conceptual disorganization, 404
- Configurational approach, 397-399, 411
  - in environment, 34
- Constructs (and criteria), 268
- Contaminated responses, 206, 415
- Content analysis, 337, 343, 396
  - related to hostility, 337
- Content categories, 120

- Convergent phenomena, 26
- Correlation, 377, 378
  - with external criteria, 216
- Creative ability, 336, 405
- Details, normal, 396
- Determinants, 13
  - common location, 190
  - different location, 191
- Developmental studies, 62
- Diagnosis, 70
  - blind, 400-405
  - compared to Rorschach, 217
  - differential, 400
- Differences, treatment of, 367
- Dimension, frequency of use of, 257
- Discrete individual event, 37
- Distribution of Rorschach signs, 206
- Divergent phenomena, 25
- Dr per cent, 210
- Dream interpretation, 407
- Drugs and personality, 341
- Electric shock convulsions, 404
- Emotionality, 120
- Empirical sign studies, 223-225
- Equivalence, coefficient of, 10
- Errors in Rorschach research, 381
- Errors, types of, 7
- Examiner effects, 134-137, 401
- Experience type, 111
- Experimental sets, influence of, 394
- Factor analysis, 408-409
  - of content, 338
- Faculty psychology, 392
- Fake Rorschach, 102
- Familial defectives, 405
- Fantasy and movement, 398
- Festinger method, 362
- Field concept, 37
- Field theory, 396
- Finger painting, 47, 67, 404
- Form-color, factorial composition of, 173
- Form-color, similarity to CF, 170
- Form level, 346
- Form, symbolism of, 407
- Free association, 56
- Frustration studies, 73
- Functional conditions, 335
- Functioning, control of, 257
- F+ per cent, 208
  - to differentiate psychotics, 208
  - and emotional control, 21
- "g" factor, 396
- Galvanic skin response, 403
- Gestalt psychology, 58, 392, 393
- Global theory, 54-56
- Ground, use of, as figure, 122
- Group method, 410-412
- Group Rorschach and MMPI, 407
- Harrower-Erickson check list, 150, 177, 408
- Homosexuals, 396-407
- Hue-incogruity, 404
- Human figure responses, 407
- Human movement, *see* movement
- Hypnosis, 203
  - and mood changes, 405
  - and Rorschach validation, 221
- Hypnotic changes, 402
- Hysterics, responses of, 396
- Idiographic, 65, 82
- Independent impressions compared with Rorschach, 216
- Individual differences, 35, 85
- Individual Record Blank, 124
- Individuation, 32
- Industrial selection, 413
- Inspection technique, 76, 409
- Insulin treatment, 409
- Integrative activity, 21
- Intellectual factors and experimental set, 394
- Intelligence, 12, 23
- Internal consistency
  - coefficient of, 10
  - criterion of, 79
- Interpersonal equation, 401
- Interpretive set and needs, 22
- Interpretive systems, 43, 68-70
- Item analysis, 236, 408
- Item selection, 5
- Language analysis, 64-65
- Life space, 38-44
- Luria technique, 217
- M responses (human movement), 163, 404-408
  - and color responses, 173
  - factorial composition of, 173
  - as a function of stimulus, 195-205
  - functional similarities among, 199
  - importance of content in, 195
  - interrelations among, 180-186
  - mutual consistency of, 181
  - See also* movement responses
- Matching method, 233-234, 235, 349, 402
- Mean, disadvantages of, 360
- Mean rank, comparison of, 362
- Measurement, 7
- Median, use of, 361
- Mental defectives, 405
- Mental life, genetic determinants of, 56
- Metrazol therapy, 223, 404-409
- Mosaic test, 59
- Movement responses, 12, 407-408
  - and day dreaming, 398
  - distinguished from C, 163
  - and fantasy, 398
  - and frontal ablations, 404
  - and self-preoccupation, 398
- Multiple choice method, 410
- Multiple regression approach, 376
- Munroe check list, 415
  - See also* check list
- Myokinetic technique, 68
- Needs, influence of, on perceptual tasks, 22
- Neurotic involvement, 406



- Neurotic phantasy, compensatory nature of, 61
- Neurotic subjects, 414
- Nomological net, 276
- Nomothetic, 65-66, 82
- Non-achieving students, 413
- Nonsense syllables, 68
- Normal detail, 84
- Normal phantasy, 61
- Normal Rorschach records, 397
- Normalizing distributions, 362
- Normals, differences among, 336
- Norms, 35, 36, 397, 398
- N-technique, 274
- Object situations, 41
- Objective data, 217
- Objective tests, 5
- Obsessive-compulsive, 399
- Officer selection, 413-414
- Operational definition, 60
- Organic differentiated from functional, 335
- Organic impairment, 406
- Organization, mechanisms of, 43, 400
- Organizational factors, 396
- Parallel series, 401
- Parent interviews, 407
- Pattern tabulation, 374
- Per cent F+, 399
- Perceptanalysis, 415
- Perception
  - of ambiguous stimuli, 104
  - of clinicians, 21
  - defined, 19, 24
  - factors influencing, 393
  - influences of, 22
  - organization of, 59
  - and personality, 14
  - and responsiveness, 19
  - of Rorschach task, 314
  - stability of, 315
  - temporal characteristics of, 393
- Perceptual control, 164-165, 197
- Perceptual task
  - classical, 19
  - and personality, 17
  - for prediction, 25
  - and Rorschach, 20, 342
- Peripheralists, 57
- Personalistic psychology, 392
- Personality
  - of artists, 413
  - blind analysis of, 407
  - definition of, 58
  - diagnosis of, 226
  - and drugs, 341
  - and hypnosis, 341
  - and per cent F+, 399
  - as a process, 34, 48
  - and Rorschach task, 22
  - of scientists, 413
  - and shock, 341
- Personologists, 82
- Phantasy, 60, 61
- Physiognomic responses, 396
- Piotrowski signs, 226
- Play configuration, 45-46
- Play technique, 45, 56
- Popular responses, 12, 122
  - related to age levels, 396
- Position responses, 206, 415
- Prediction
  - methods of, 412
  - and probability, 3
  - from Rorschach task, 216, 409
- Probability, 3, 4, 354
- Projection
  - of individual personality, 43
  - versus projective, 55
- Projective data and case study, 216
- Projective methods, 43-44, 53
  - classified, 67
  - definition of, 53
  - purposes of, 70
  - use of, 68
- Projective test situation, 19
- Projective tests, 5-6
- Proportions, tests for significance of differences of, 352
- Psychodrama, 68
- Psychometric scales, 18, 396
- Psychoneurotics, Rorschach signs for, 203
- Psychopathology, 62
- Psychotherapy, 71
  - behavior in, 407
  - Rorschach changes in, 338
  - Rorschach signs for, 203
  - and Rorschach task, 398
- Q technique, 419
- Qualification, problems of, 77
- Ratings
  - disability, 406
  - efficiency, 413
  - instability, 406
  - teachers', 407
- Reductive effort, 58
- Regression by hypnosis, 103
- Relationships, subject-examiner, 401
- Reliability, 35, 77, 297, 377, 401, 409
  - assumptions for interpreting, 10-11
  - assumptions with Rorschach test, 314
  - defined, 10
  - lack of, in Rorschach test, 333
  - problems of, 296
  - and validity of tests, 41
  - and various length records, 379
- Reliability, types of, 310
  - independent estimates, 379
  - parallel series, 297
  - split half, 298, 402
  - test-retest, 297
- Research, projective techniques in, 72
- Responsiveness, selective, 34
- Roles, 47, 255
- Rorschach administration
  - group method, 409
  - multiple choice, 414
  - negatively loaded, 106
  - positively loaded, 107
  - varied instructions, 394
- Rorschach as an experiment, 342

- Rorschach assumptions, 18, 147, 342, 403
- Rorschach categories
  - changes with administration, 129-131
  - comparing groups by, 358
  - and diagnosis, 335
  - factorial composition of, 171
  - and independent measures of personality, 228
  - and interpretation, 174
  - patterns of, 372
- Rorschach productivity, 320
- Rorschach Ranking test, 414
- Rorschach, related to
  - case history, 399
  - drugs, 221
  - hypnosis, 220
  - MMPI, 407
  - Wechsler-Bellevue, 407
- Rorschach responses
  - distribution of, 417
  - functional similarity of, 148, 154, 189
  - intercategory differences, 194
  - intracategory similarities, 194
  - patterns of, 101
- Rorschach signs
  - associative elements of, 19
  - in brain damage, 203
  - comparison of, 373
  - and psychiatric groups, 204
- Rorschach statistical considerations
  - clinical validity, 334
  - continuous scale, ratings on, 349
  - dichotomized ratings, 348
  - inequality of units in scales, 360
  - objectification, 394
  - standardization, 394
- Sampling, 85
- Schizophrenia
  - patients in psychotherapy, 409
  - process, 406
  - records, 403
  - signs, 203
- Scoring
  - absolute, 7
  - agreement, 311
  - categories, 21-23
  - factor analysis of, 338
  - objective, 333
  - quantification, 179
  - summary of, 203
- Selection
  - attentive, 59
  - mechanisms of, 400
- Self-adjustment, 400
- Self-description, 407
- Sensation, 19
  - and cognition, 22
- Sex populars, 396
  - responses, 210
- Signs, 415
  - and sample tests, 6-7
- Single case research, 66
- Situational factors, 103
  - influences, 394, 407
- Skewness, 368
- Small samples, significance tests for, 351
- Socialization, 32-35
- Sodium amytal, 103, 405
- Spatial-temporal configurations, 42
- Standardization, 35, 77-80
- Statistical approach, 3
  - studies, errors in, 351
- Stimulus
  - ambiguity, 24
  - apart from field, 41
  - external, 19
  - fallacy, 69
  - projective, 25
  - role of, 339
  - situation, 43
  - value of, 5
- Structure, internal, 273
- Stutterers, 405
- Subjective factors, 394
- Suicidal trends, 409-412
- Symbolism, 224
- Tachistoscopic exposure, 404
- Task, influence of repetition on, 133
- Tautophone, 59, 68, 83
- Test situation, subject's definition of, 103
- Tests, types of, 5
- Theory construction, 23, 83-84
- Time per response, 116
- Topectomy, 338
- Topological concepts, 56
- Transference, 104
- Treatability, determination of, 409
- Type, dimensions of, 255
- Typology, 392
- Unconscious complexes, theory of, 63
- Validating procedures, 228
- Validation, 216, 221
  - clinical, 78
  - criterion oriented, 265
  - item by item, 233
  - subjective, 333
  - temporal, 40
  - and test inference, 285
  - of tests, 285
  - through prediction, 216
  - time consistencies, 216
  - types of, studies, 233
- Validity, 34, 78, 403-405
  - blind interpretation, 76
  - and constructs, 272-275
  - defined, 8, 218
  - external, 11
  - and factor analysis, 275
  - logic of, 276
  - and psychiatric diagnosis, 76
  - studied indirectly, 403
  - types of, 8-9, 264-265
- Verbalizations, deviant, 338
- Visuomotor conflict, 403
- Vocational guidance, 62, 410, 413
- W responses, 399
  - and organization, 20
- Word association, 63

